Evidence-based Education: The development of a model to use protocols and small-scale aggregated trials to create a prospective cumulative meta-analysis as an evidence base for interventions.

HARRISON, WAYNE

Evidence-based Education: The development of a model to use protocols and small-scale aggregated trials to create a prospective cumulative meta-analysis as an evidence base for interventions.

Wayne Harrison

Doctorate of Philosophy in Education

School of Education

Durham University

February 2020

To:

Angela and Amy

Acknowledgement

I would like to start by thanking my wife and daughter for their support and understanding while completing this thesis. Without your understanding and patience over the late nights and weekends while I combined studying with creating a business, this piece of work would not have been possible. I would also like to thank my parents for their support when I decided to become the first in my family to study A levels and then attend university, as without the experiences of my teaching career this study would not be possible. My acknowledgement must also thank Prof Steve Higgins, firstly for providing the support in changing from a teaching career and securing a funded ESRC for an MA and PhD at Durham University. Secondly for the understanding and many conversations on broader areas of education, as these have helped to shape my understanding of evidence-based education. I would also like to thank Prof Christine Merrell and Prof Carole Torgerson for their invaluable feedback while developing this thesis. Dr ZhiMin Xiao, Dr Akansha Singh and Dr Germaine Uwimpuhwe for their understanding and patience for the advice on the statistical analysis and Sean Gardner for allowing access to the online platform for the primary research studies. The research would not be possible without all the schools, teachers and pupils who have taken part in the studies and for the ESRC support in funding the MA and PhD at Durham University. Thank you to everyone for your support over the last five years.

Abstract

In education, there has been a worldwide increase in the use of evidence in education to inform policy and practice. In the USA, bodies such as the Institute of Educational Sciences' (IES) What Works Clearinghouse (WWC), the Best Evidence Encyclopaedia (BEE) and the Department General Administrative Regulations (EDGAR) and in the UK the Education Endowment Foundation (EEF) have been established to inform education decision making. The global increase in the use of evidence in education is based on the premise that if programmes are selected on the basis of more robust evidence, using these tested interventions should increase the chance of positive outcomes if they are deployed in other schools and contexts.

The aim of this thesis is to explore evidence-based education in the literature review before proposing a new theoretical model for evidence generation, introducing the use of protocols and small-scale aggregated trials linked to a prospective cumulative meta-analysis (PCM). The application of a PCM allows replication as a way to test how likely the intervention effect sizes will translate when they are tested using large scale randomised controlled trials and also as a method to test stability and improve dissemination after the initial research. The thesis includes two primary research studies for online cross-age peer tutoring across the transition boundary between primary and secondary schools and online small group teaching. The purpose of the trials were to test the implementation of the methodology for the use of protocols and small-scale aggregated trials linked to a prospective cumulative meta-analysis (PCM). Study One focuses on online peer tutoring across the transition for primary and secondary pupils and Study Two investigates the effectiveness of online small group teaching for mathematics.

The thesis demonstrates how the use of the model can be used to increase replication in the testing phase of an intervention, using empirical evidence from the online peer tutoring trials involving data collected across three cohorts of schools in an academic year. The impact of this research will provide an alternative testing framework for deciding in future trials if the evidence is robust for the commissioning of large scale randomised controlled trials in education.

Contents

Figures

Tables

List of Abbreviations

AR – Accelerator Reading Programme

BEE – Best Evidence Encyclopaedia

CG – Campbell Group

CONSORT – Consolidated Standards of Reporting Trials

EDGAR - Department General Administrative Regulations

EEF – Education Endowment Foundation

ESRC – Economic and Social Research Centre

ITT – Intention to Treat

KMb – Knowledge Mobilisation

NICE – National Institute to Health and Care Excellence

PCM – Prospective Cumulative Meta-analysis

REML – Restricted Maximum Likelihood

RCT – Randomised Controlled Trial

SPIRIT – Standard Protocol Items: Recommendations for Interventional Trials

TTA – Teacher Training Agency

TUTE – Online learning platform

WWC – What Works Clearinghouse

1 Introduction

1.1 Context in the UK

In education, there has been a worldwide increase in the use of evidence in education to inform policy and practice. In the UK, the EPPI Centre and Education Endowment Foundation (EEF) have been established with the purpose of improving the evidence for practitioners to use to inform policy and practice. The EPPI centre is based in the Institute of Education at University College London. It specialises in developing methods for systematic reviewing and synthesis of evidence, as well as developing methods for the study of the use of research. The EEF is an independent grant-making charity dedicated to breaking the link between family income and educational achievement. The charity was founded by the Sutton Trust in partnership with Impetus Trust, with a grant of £125 million from the Department of Education. The charity funds and independently evaluates educational interventions, investing in evidence-based projects designed at tackling the attainment gap between disadvantaged students and their peers. The creation of the Sutton Trust–EEF Teaching and Learning Toolkit supports teachers, school leaders and policy makers to make informed decisions about practice based on robust evidence. The development of the Teaching and Learning Toolkit (Higgins et al., 2015) is specifically designed to evaluate interventions, through large scale RCTs and meta-analyses of existing evidence.

Over the past two decades there has been increasing interest in and focus on using Randomised Controlled Trials (RCTs) as the best way to evaluate educational interventions, often labelling these the 'gold standard' for evaluations (Ginsburg & Smith, 2016). The increase in the use of systematic reviews and meta-analyses in education has prompted more rigorous approach to evidence collection, combined with an increase in experimental studies to develop an evidence base for practitioners. Yet, as stated by Rothstein et al. (2005), confidence in meta-analytic and systematic reviews is dependent upon the extent to which these findings are accurate and free from bias. If the literature from the evidence-base is used to inform the design of educational interventions, and the initial evidence is flawed, we have a fundamental problem at the start of the research cycle as even the most robust RCTs will provide uninformative evidence.

Furthermore, recent studies (Cheung & Slavin, 2016; Ginsburg & Smith, 2016) have raised concerns about the usefulness of a number of RCTs and meta-analyses. The authors have no quarrel with the logic and potential methodology of RCTs in education, as they agree that a well implemented RCT is an important tool for establishing an unbiased estimate of a causal effect for an educational intervention. The strong internal validity of RCTs does have a potential drawback, as they are often

carried out with a particular sample, at a specific time, particular place and in a specific setting. Therefore, the external validity of RCTs can be a potential issue. The issue is compounded by a 'replication crisis' in educational research and also mirrored in psychology research. Makel & Plucker (2014) found that only 0.13% of articles in leading educational research journals are replicated studies, which increases the risk of flawed research being used as a basis for the foundation of an intervention.

Consequently, if interventions are designed based on unreliable research these are unlikely to be effective, even if the design of the study is excellent and the intervention is implemented correctly. This is important ethically as time and resources are required to robustly evaluate interventions. For example, the EEF spends approximately £500,000 per trial (EEF, 2015) and a recent study by Lortie-Forgues & Inglis (2019) found that the average effect size from high quality RCTs was 0.06 standard deviations. Therefore, before committing to scale an intervention for a rigorous RCT, we need to develop a methodological approach to allow small scale replication of studies and use this evidence to inform whether we are likely to replicate the positive effect size at scale. Furthermore, it can be argued that it is even less ethical to roll out an intervention with any form of evaluation, as the government has done in the past by introducing new policies without careful evaluations.

The purpose of this thesis is to review the literature on evidence-based education and suggest a model for using protocols and small scale RCTs to create a prospective cumulative meta-analysis (PCM). The development of a PCM creates an evidence base to inform if an intervention should be moved to the rigorous testing phase of the research cycle. The thesis involves two primary feasibility studies, replicating small scale RCTs using a protocol and generating a PCM to demonstrate how this methodology can improve the testing phase of the research cycle proposed by Gorard (2013). The first primary study is a replication of the methodology used in an ESRC funded MA at Durham University (Harrison, 2015), investigating the effectiveness of online peer tutoring group sizes for an intervention used to support academic transition between primary and secondary schools. The study builds on the previous research by replicating a pilot with two further cohorts of schools using the same protocol, allowing the data to be aggregated in a PCM. The second study evaluates the effectiveness of online small group teaching using a similar methodology. It is important to acknowledge that these studies are only intended to generate evidence for the proposed model; due to the small sample sizes no generalisations are intended for the effectiveness of either intervention.

To conclude this research is important as it is imperative that the evidence within education should be of the highest quality, accurate and free from bias to enable practitioners and policy makers an evidence base which they can use to inform their decisions. Through the development of a new testing and stability phase of the research cycle, we can better inform whether an educational intervention is likely to have a positive impact in real life settings such as schools.

1.2 The structure of this thesis

Following this short introduction, the thesis is divided in three main parts. The first part covers a literature review of evidence-based education, the second part includes the primary research studies and the third part covers the development of protocols, methodological advantages of using a prospective cumulative meta-analysis and finally reflections and discussions of the findings.

1.2.1 Part One

Chapter 2 sets out an introduction into the evidence-based movement and explores the debate in the literature for the use of evidence in education. The chapter progresses onto the meaning of evidence, defining evidence-based education and how this links to the global movement in this field of research. The importance of the research cycle in social sciences is introduced and the research evidence and conceptual frameworks discussed on the strategies used to increase the use of research in education.

The chapter moves onto how prospective cumulative meta-analyses are used in healthcare to aggregate data from a number of studies before proposing an introductory model for the social sciences. The model allows the use of protocols to replicate small scale RCTs and aggregates data into a prospective cumulative meta-analysis (PCM) as a method to increase replication and provide a robust evidence base for educational interventions.

1.2.2 Part Two

Chapter 3 provides an overview of the two primary studies involved in the development of small-scale aggregated trials linked to a prospective cumulative meta-analysis. The first section outlines the logic of appropriateness for the use of randomised controlled trials as the preferred methodology, specifically linking to internal and external validity. The logic of appropriateness of the research design is discussed on why RCT methodology is used in each primary study.

Chapter 4 involves the first primary research methodology for the pilot study for small scale aggregated trials investigating the effectiveness of online cross-age peer tuition group sizes across the transition boundary between primary and secondary schools. The research follows on from an initial feasibility study (Harrison, 2015) into the use of online cross-age peer tuition in mathematics between primary and secondary pupils in 4 UK schools.

Chapter 5 presents the impact and process evaluation for the online peer tuition study investigating the effectiveness of the group sizes 1:2 and 1:4 compared to the 1:1 active control. The three

cohorts of schools enabled the data to be aggregated into a prospective cumulative meta-analysis, the first instance of this methodology used in educational research, so far as I am aware.

Chapter 6 outlines the second primary research methodology for a feasibility study to develop a prospective cumulative meta-analysis linked to small scale RCTs for the effectiveness of online small group teaching using the TUTE mathematics programme.

Chapter 7 presents the impact and process evaluation for Study Two. It was not possible to produce a PCM due to only one cohort of schools completing the research, however the study provides further evidence for the potential for using online small group tuition as an educational intervention.

1.2.3 Part Three

The final part of the thesis starts with Chapter 8, exploring the lessons learned in creating trial protocols in healthcare and reflecting on the lessons learned from clinical trials for the development of the SPIRIT 2013 guidelines. The chapter proposes a modified version of the SPIRIT guidance for the social sciences, before discussing how implementing these guidelines will increase the transparency of the evidence-base in education.

Chapter 9 discusses common methodological issues regarding the quality of evidence in educational research, such as the type of publication (published versus unpublished), size of sample (small, N< 250 versus large, N > 250), research design (randomised versus matched) and outcome measurements (experimenter made vs independent). The chapter presents an argument for the use of prospective cumulative meta-analyses linked small scale aggregated trials as a method which can reduce or eliminate these concerns.

Chapter 10 discusses the findings from the primary research studies, lessons learned for implementing small scale RCTs in the development of a prospective cumulative meta-analysis and makes the case for why social sciences should adopt a standardised approach to evidence-based education. A proposed model is presented for using aggregated trials linked to a PCM in evidence-based education. The chapter then progresses on to review the study limitations; discussing these in terms of theoretical, methodological and implementation issues before suggesting key recommendations for policy and practice. Finally, the conclusion outlines the importance of the thesis to the current evidence base in education and why we need to increase transparency across all forms of educational research, not just in studies involving experimental designs.

Chapter 2: Literature review

This chapter will introduce the evidence-based movement in education, causality and a comparison of counterfactuals in research designs. Secondly, the social science research cycle will be outlined, highlighting a potential flaw in the current research cycle. Thirdly, models of evidence-based practice and then the conceptual models for dissemination will be discussed. Fourthly, a review of the empirical literature about the use of research by educational practitioners and the factors affecting this use will be critically analysed. The chapter concludes by exploring an alternative evidence base using school led aggregated RCTs and linked to a prospective cumulative meta-analysis (PCM) to create testing and stability phase of the research cycle.

2.1 Introduction to the evidence-based movement

The idea that teaching should become an evidence-based profession has recently come to prominence in a number of countries over the last few decades. However, in the United Kingdom the issue regarding the state of educational research came to light following Professor David Hargreaves comments in his 1996 Teacher Training Agency (TTA) lecture entitled "Teaching as a research-based profession: possibilities and prospects" (Hargreaves, 1996). In the lecture, Hargreaves raised a number of concerns regarding the current state of educational research which resulted in a lively debate among educational researchers.

Firstly, educational research was poor value for money in terms of improving the quality of education provided in schools, with an annual spend of £50-60 million (Hargreaves, 1996). Secondly, he argued that educational research was non-cumulative as few researchers seek to create a body of knowledge that can be tested, extended and replaced using systematic approaches. The comparison with medical research and the cumulative nature highlights the issue in education of the unsystematic approach to research and use by teachers. Thirdly, Hargreaves takes issue at the lack of relevance for research in improving classroom practice, with academics engaged in methodological quarrels which are pointless to anyone outside of the academic communities. The final issue related to the fatal flaw in educational research, the gap between researchers and practitioners. Hargreaves explains that in medicine, there is little difference between researchers and users, as all are practitioners. He contrasts this with education, whereby researchers are rarely users, creating issues with communication and dissemination of findings. Within medical research, enabling practitioners to become enquirers using scientific methodology as part of any enquiry also improves the level of quality across the profession.

As previously stated, these criticisms provoked a debate (Bassey, 1996; Gray et al., 1997; Hammersley, 1997) and provided the impetus for a number of government funded publications (Bassey & Constable, 1997; Hargreaves, 1996; Tooley & Darby, 1998). The essence of the debate centred around whether the evidence could be produced similar to medicine. Both sides shared the commitment to quality in educational research and teaching, with Hammersley (1997) arguing that 'Hargreaves seems to assume a strong form of instrumental knowledge at odds with the practical nature of teaching'. In response to the concerns raised, Bassey & Constable (1997) analysed 12,000 educational research papers submitted to the Education Panel for the Higher Education Research Council 1996 research assessment exercise. The titles were screened and the articles placed into eight categories; curriculum (30%), school/teacher/child (11%), teaching/learning (7%), governance (4%), phase areas (23%), overseas studies (7%), disciplines (7%) and methodology (1%). The authors concluded that the concerns about educational research were misplaced, as curriculum accounted for almost a third of the research and another one fifth focusing on the school/teacher/child and teaching/learning. However, a flaw in the research existed as the study only screened titles of the research articles with no reference to the quality of the research. In hindsight, this report encapsulates the issues regarding rigour with a large number of educational research in the late 1990's and early 2000's.

Tooley & Darby (1998) acknowledge the report created by Bassey & Constable (1997) as a useful mapping exercise of the research terrain, although the figures only represent the areas of research and explain nothing about the quality of the literature surveyed. OFSTED (Office for Standards in Education) commissioned Tooley & Darby (1998) to assess the quality of papers in the four top-ranking educational research journals according to the SSCI's Journal Impact list. The initial sample included 264 articles, with a sub-sample selected and reduced to 41 to ensure the coding for the 30 research questions was manageable and allowed a snapshot of the quality of each journal. The analysis found four major themes: the partisan researcher, methodological problems (such as sample size and selection), non-empirical research and the focus of educational research. In addition to these themes, 26 articles were highlighted as not satisfying the criteria for good research compared to 15 articles classified by the authors as good practice. The authors raised questions regarding the process of peer review, as all the papers included had been accepted through the academic referring process and discussed poor state of the educational research community.

In response to the Ofsted commissioned report, a number of academics raised objections to the detailed critique (Hextall et al., 1998; Vulliamy, 1998) by Tooley & Darby, who take issue with the interpretation of their research and writing. Furthermore, they argued a lack of understanding regarding the complexity of qualitative research and the criteria used to assess the quality

demonstrated a bias towards the research (Hodkinson, 1998). Consequently, the fierce defence by a number of academics to the published report highlights the issues raised by Hargreaves (1996), as the concerns regarding sampling bias, selection bias, sample sizes and generalisability (external validity) were genuine methodological concerns and cannot be defended through a lack of understanding of the complexity of qualitative research.

In response to Hargreaves' TTA lecture, Hammersley (1997) published a detailed response critiquing a number of the issues raised. The main criticisms involve the comparison of educational research with medicine as potentially misleading. Hammersley argues that "medical research does not involve the distinctive problems associated with social phenomena" and financially, medical research attracts significantly higher levels of funding than educational research. Additionally, through the practical character of teaching, the diverse and difficult to operationalise goals, complex relationships and multiple variables can mean that educational research can seldom provide solid information regarding effectiveness. These critiques were defended in a rejoinder published by Hargreaves (Hargreaves, 1997) and resulted in a number of debates within the educational academic community on the quality and status of educational research.

2.2 The debate on evidence-based education

It is not possible within the limitations of this chapter to outline the history of the arguments relating to evidence-based education, yet a number of recurring themes appear in the literature. The proceeding section will outline the criticisms and counter arguments for evidence-based education.

Argument 1: It is not possible to compare educational and medical research

A number of academics have argued there are problems associated with measurement and attributing causation in education, which are not found in medical research (Gray, 1996; Hammersley, 1997; Morrison, 2001; Norris, 1996). These academics thus claim that education cannot be compared to healthcare, more specifically to medicine, because educational research involves social interactions, with processes and outcomes which are complex, often culturally and context specific.

Davies (1999) counters this argument by explaining that medicine and healthcare face very similar problems relating to complexity, measurement, context-specificity and causation. These issues have been raised in the medical literature, as well as the generalisability of evidence-based healthcare. Davies provides specific examples of these concerns and parallels with education, such as the ecological validity of the uncertain relationship between how people behave in hospitals and other environments such as their own homes (Andrews & Stewart, 1979).

Argument 2: RCTs are an inappropriate methodological procedure for educational research

A recurring critique appears in the literature (Biesta, 2007; Hammersley, 1997; Morrison, 2001) with regards to the methodology of RCTs, with these authors arguing that it is not possible to identify all the variables under experimental conditions in real life settings such as schools. To then suggest that these variables could be isolated and controlled to create equivalent groups is not possible, therefore preventing the researcher attributing causality. Morrison (2001) disputes the ability of RCTs to meet the requirements of the principles on which they are based, making RCTs fallible. Furthermore, Biesta (2007) suggests that treating education as an intervention or treatment as a causal means to bring about pre-established outcomes is not appropriate in education. He implies that the methodology of RCTs is not appropriate, as the most important questions for educators are not about the effectiveness of their actions but about the potential educational value that they provide to students. Thereupon, Biesta acknowledges that education is a moral practice, rather than a scientific or technical practice.

Coe, Fitz-Gibbon, & Tymms (2000) contest that RCTs and systematic reviews are the most appropriate approach for evaluating educational interventions designed to change policy, practice and educational systems. It is only through robust evaluations that test a counterfactual comparison that causation can be attributed to interventions. Fitz-Gibbon (2004) provides examples that illustrate that guessing and good intentions in education are not the basis for effective action. The appropriateness of the methodological advantages of RCTs compared to other research designs are raised by Slavin (2002), who explains that when selection bias is a possibility at the student level, there are few alternatives to the process of randomisation, as often unmeasured pre-existing differences are highly likely to be alternative plausible explanations.

Argument 3: The use of RCTs in education is not ethical

Torgerson & Torgerson (2001) highlight that the methodology of RCTs has been criticised as being unethical because an intervention is withheld from some children. Yet, the purpose of the trial is to determine if the intervention is effective and if the researchers knew that the intervention was effective this would indeed be unethical. However, this is not the case. Gorard (2013) explains that it is possible to ensure that an intervention can be evaluated while offering the treatment to all cases using a wait-list design. In this instance, all cases are offered the treatment, but they are randomly allocated to receive this at the start of the trial or after a delay. Therefore, the ethical concerns raised are usually eliminated whilst enabling researchers the opportunity to determine if the intervention works or does not.

Moreover, the question must be asked if it is unethical to not conduct a trial as the consequences of implementing interventions which are not rigorously evaluated can potentially damage the educational outcomes of students, whilst wasting time and finances.

Argument 4: The democratic deficit of the promotion of evidence-based education

Biesta, (2007) examined the model of professional action that is implied in the concept of evidence-based education, arguing that we should not understand education as an intervention or treatment. Biesta states that he believes education to be non-casual and the normative nature of practice, which implies the need for educational practitioners to make judgements about what is educationally desirable. Biesta argues that replacing practitioner professional judgement with "what works" evidence, creating a democratic deficit in evidence-based education whereby teacher judgements become limited and hence devaluing educational decision making. Similarly, Hammersley (1997) suggests that evidence-based medicine could replace clinical judgement and threatening professional power in the healthcare profession. He draws parallels with the education profession and the potential threat to teacher professional judgements.

Higgins (2017) responds to these criticisms as creating a false dichotomy, as he believes it is possible to hold a scientific view of causation at the same time as considering education as a process of symbolically mediated interactions. Higgins explains that through using macro and micro trials, it is possible to find out if an intervention is effective in terms of the outcomes in the trial conditions. The high internal validity of well-conducted RCTs, if the sample is sufficient, allows causation to be inferred for the sample included in the trial. However, further professional judgement is required by practitioners to decide if the intervention will be effective in their school and for their pupils. Through thinking about the distributions around the average treatment effects, rather than a definitive effect size the evidence-based education model creates a 'best bet' judgement for professionals. Surely allowing informed decisions based on evidence creates a more democratic gain rather than deficit for educational professionals?

2.3 What do we mean by evidence-based education?

It is important to state before defining the term evidence-based education, that educational research encapsulates a wide range of research designs and methodologies. It is not the intention to discount the merits of ethnographic studies, case studies, longitudinal studies or any other type of research as these may be relevant to practitioners in some way. Yet, the concept of evidenced-based education relies on the premise that a key component in educational research is to bring about intentional change in educational outcomes of students, educational systems and practitioners. If the intention is to assess the impact of these interventions, then the research questions underpinning the evaluation are predominantly causal in nature.

Furthermore, a subtle difference is evident between the terms evidence-based education and evidence-informed education and between evidence-based practice and evidence-informed practice (see Nelson and Campbell, 2017 and the articles they introduce for a fuller discussion of this debate) . These terms are contested, and the role of the practitioner in generating and using evidence is at the centre of this debate (Cordingley, 2008). The development of the use of protocols and small-scale RCTs for practitioners in this study is to generate evidence which aggregates into a prospective cumulative meta-analysis implies that evidence is created to build the evidence base from the ground up, from practitioners experiences in classrooms. Evidence informed suggests a top down approach, with research created and consumed by practitioners, though supported and interpreted by professional expertise (Hedges, 2012). A study commissioned by the EEF for teacher choices aims to address concerns from teachers and headteachers that research is not linked to teacher practice, so this study directly feeds into existing teaching practices (Styles, Galvis, Lord et al., 2019). The model proposed in this thesis is different, as the small-scale studies allow replication and cumulation of data over a sequential time period, rather than a fixed model with the direct involvement of teachers as professionals. Additionally, the use of a protocol with the resources and the assessments standardised across each research replication should reduce the heterogeneity within any aggregation of findings and make it easier to draw more robust conclusions about generalisability.

Therefore, before attempting to define what is meant by the term evidence-based education, it is important to discuss the concept of causation in relation to impact evaluations. According to Mill (1882), three conditions are required to infer a causal claim. Firstly, the cause preceded the effect. Secondly, the cause and effect are related. Thirdly, there is no plausible alternative explanations for the effect other than the cause. However, the criteria set out by Mill fail to explain what is meant by the term related and this develops uncertainty in inferring causation.

Bradford-Hill (1966) and Cook & Campbell (1979) proposed criteria for establishing causation and expanded on Mills work by adding a requirement for varying the strength of a cause which results in a change to the effect. However, as Gorard (2013) correctly states, these criteria fail to include the crucial element for the elimination of sensible alternative explanations for the change. In order to establish causation, Gorard (2013) identifies the standard ingredients for a casual claim as a correlation between a cause and effect; the effect appearing after the cause; varying the strength of the cause changes the effect produced and it is a plausible explanation of the process by which the change happens. In terms of impact evaluations, if the intervention (X) and the possible effect such as an improvement in educational outcomes (Y) has a causal relationship, then they must be repeatedly associated, replicable, observable and specific to X and Y. Additionally, X must always precede Y when they both appear, and the addition of X can predict the appearance of Y. By changing the strength of X it should change the strength of Y and the casual mechanism must be explained at its simplest form.

Therefore, in order to establish if an intervention (X) caused the effect of increasing attainment in Mathematics (Y), the researcher must ensure that no other plausible explanations can explain the increase in mathematics attainment. The following section will suggest a definition for evidence-based education, counterfactual comparisons linked to research design and discuss the implications for creating a robust evidence base in education.

2.3.1 Defining evidence-based practice

Coe, Fitz-Gibbon, & Tymms, (2000) define 'Evidence-Based Education' as "the support and promotion of practices and policies that are based on good evidence about their effects (i.e. costs and benefits)". The implication is that evidence-based education involves interventions, with actions such as providing advice, implementing policies and promoting or requiring a specific practice are all classed under the term intervention. Consequently, to determine if an intervention is successful a change must be made, rather than simply observing or describing existing situations. Furthermore, the benefits of the intervention must be greater than the cost of implementing the change. For this reason, it is imperative that the intervention is evaluated robustly and under effectiveness conditions in educational settings.

The definition appears to imply a hierarchy through the inclusion of the term 'good evidence', whereby the type of research design implies a level of confidence in that the change or intervention was actually responsible for any changes in outcomes. The research questions in evaluations of

interventions are usually causal in nature, as policy makers, researchers and practitioners want to find out if the intervention was responsible for any improvements.

When considering the effectiveness of an educational intervention, whether this be policy or practice, it is important to understand the counterfactual or the comparison being made. The counterfactual is crucial, as it is impossible for a student to experience the intervention or not at the same moment in time. Consequently, the counterfactual cannot be directly observed and this must be compared to a comparison group. Higgins (2017) suggests that the different kinds of comparison, through different research designs, provide evidence for a stronger or weaker argument about the robustness of any causal claims regarding the effectiveness of an intervention. Furthermore, Higgins (2017) states that the strength of a claim weakens as the number of threats to the internal validity increases, meaning that the counterfactual comparison becomes less convincing. Table 1 shows a comparison of research designs used in evaluative research designs and their counterfactual comparisons, with the strength of inferences reducing as you move down the table.

Table 1: A comparison of the counterfactuals and threats to the internal validity in evaluative research designs (Adapted from Higgins 2017)

	Design	Counterfactual	Internal validity
Experimental	Randomised controlled trial	Compares the average outcomes from groups allocated by random assignment to treatment with a group that does not experience the change, creating equivalent groups for unknown and known variables with a sufficient sample size.	Very strong internal validity as RCTs control for selection bias, temporal changes, regression to the mean and instrumentation threats. Three armed trials can control for the effects of innovation, using 'business as usual' with a 'placebo' comparison. Limitations: Attrition and social interaction threats can affect internal validity.
	Randomised regression discontinuity	Students can be randomised around a cut off score using a statistical model. Compares students around the cut-off with all student outcomes in the sample.	Randomisation around a cut off score controls for selection bias and maturation effects. Limitations: Randomised regression discontinuity does not control for effects of innovation or novelty. It has an underlying assumption that the pre-post relationship can be accurately modelled.
	Quasi-experimental design	Provides a comparison of the average outcomes from students who are non-equivalent with a group that does not receive the change.	Provides a control group as a counterfactual. However, this is limited in inferring causation. Limitations: The design does not control for selection bias, temporal changes, regression to the mean effects and instrumentation threats. Unlike an RCT, this does not control for and any unknown characteristics which may influence learning outcomes. Attrition and social interaction threats are also potential issues.
Observational	Natural experiments Matched comparison groups Difference in difference	Similar students who do not experience the intervention (change) are compared with outcomes of students exposed to the change.	In order for an analysis, groups must be sufficiently similar to compare. Limitations: These do not control for selection or bias that is related to unobserved or unmatched characteristics.
	Time-series (e.g. single group design)	Compares the outcomes of the same students before and after a change.	No comparator group is used in this design, allowing minimal casual inference if the input and output variables correlate strongly. Limitations: Selection bias, regression to the means, maturation and temporal changes, instrumentation threats, attrition and social interaction threats.

To recap, evidence-based education is based on two premises. Firstly, the implementation of an intervention or change at either policy, practitioner or student level. Secondly, the intervention must have been evaluated using the most robust research design to allow confidence in inferring causation. Coe, Fitz-Gibbon, & Tymms, (2000) suggest using a systematic review of multiple RCTS of interventions conducted under effectiveness conditions in educational settings. Through the use of randomised controlled trials, the threats to internal validity are minimised compared to the other designs outlined in table 1, reducing alternative explanations for any changes in educational outcomes.

However, Biesta (2007) argues against the idea that education is a causal process as he states that education is not a process of physical interaction but a process of symbolic or symbolically mediated interaction. He argues that if teaching is to have an effect on learning, it is the students who interpret and make sense of what is being taught so it is the mutual interpretation process which makes education possible. Hammersley (1997) provides similar concerns for causality in education, as he states that problems relating to the establishment of causal patterns are equally severe as researchers are interested in what goes on in real schools and colleges. Referencing complex webs of relationships and unlike medicine, in education the 'treatments' consist of symbolic interactions with a scope for multiple interpretations and responses.

Coe, Fitz-Gibbon, & Tymms, (2000) acknowledge the argument that teaching is a complex interaction and that opponents of evidence-based education refer to the importance of context and subtle interactions between people. These interactions are interpreted and responded to by the agents, shaping the environment with an unpredictable nature. Morrison (2001) uses complexity theory to replace the emphasis of causality, stating that small scale changes in initial conditions can produce massive and unpredictable changes in outcomes. However, as Coe, Fitz-Gibbon, & Tymms, (2000) rightly point out, if we are saying that there can never be an intervention whose effects are predictably beneficial then we are effectively stating that we cannot provide advice to practitioners or policy makers. Do we let politicians and 'experts' advise teachers and schools based on their knowledge of what works? Or do we want a system whereby evidence is generated using a systematic approach to finding out what works on average, and then professional judgement brings in knowledge of an individual and their context.

To conclude this section on what is evidence-based education, it is clear that interventions in policy or practice are intended to make a positive change to the educational outcomes of students, practitioners or educational systems. Therefore, the research questions underpinning the evaluations must have a causal inference, as in these instances purely observational or descriptive research would not allow causality to be attributed. Chapter 3 focuses on the methodology of randomised controlled trials and the logic of appropriateness for addressing research questions relating to causation.

Additionally, it is important to acknowledge that poorly designed randomised controlled trials or even well designed studies can have a number of threats to their internal validity and external validity. These will be discussed throughout this thesis as an argument for the introduction of school-based small scale aggregated trials. For the purpose of this thesis, evidence-based education is therefore seen as complementary to and supportive of evidence-informed practice. Indeed, one of the key arguments of the thesis is that we need to take a more critical view of current conceptions of evidence-based education to develop a more robust and rigorous evidence base which is built by classroom practitioners (Cordingley, 2008). This is through undertaking incremental and cumulative research investigations which help us understand the reliability and replicability of the evidence in relation to a specific change.

At this point in the thesis, it is also important to acknowledge that a pragmatic approach will be taken in context of the studies involved and the development of a prospective cumulative meta-analysis. Consequently, the epistemological and ontological positions will only be covered briefly as this is not the purpose of the thesis.

Within the context of this thesis the term 'research' refers to a notion of systematic logic of enquiry in answering a particular question (McMillan & Wergin, 2006). In educational research, and the social sciences more generally, there has been a move away from the previous dogmatic attachments to positions such as "positivist", "rationalist", and "interpretative" stances. According to Denscombe, there is a growing tendency to combine research methods and he argues, 'it is not how well it sticks to its "positivist" or "interpretivist" epistemology, but how well it addresses the topic it is investigating.' (Denscombe, 2003).

Academics such as Mason (1996) disagrees against taking a pragmatic approach to research positions, believing that all research should start by figuring out their ontological and epistemological position before considering research questions and methods.

Mason argues against a pragmatic approach, believing that researchers start by figuring out their ontological and epistemological position (Mason, 1996) before considering the research questions and methods. However, it can be argued that by deciding your ontological and epistemological positions before starting your research, these may mislead or distract from the researchers' ability to have a neutral stance when defining their research questions and using the most appropriate research methods.

2.3.2 Research cycle for social sciences

In the development of an educational intervention, it should progress through several stages between its initial development and dissemination into routine practice in schools (Greenburg et al. 2005). In clinical studies, Campbell et al., (2000) outlined the sequential phases for developing complex interventions. These stages include: pre-phase to develop the theory, phase 1 involving modelling empirical relationships with the intended outcomes, phase 2 for exploratory trialling, phase 3 for efficacy RCTs under optimum conditions and finally phase 4 involving long term effectiveness studies under normal conditions.

In the US, the Institute Educational Sciences' WWC has developed a framework similar to the Campbell (2000) using research goals for funding applications. These five goals include: Exploration (Goal 1), Development and Innovation (Goal 2), Efficacy and Replication (Goal 3), Effectiveness (Goal 4) and Measurement (Goal 5).

Research in the social sciences generally follows a cycle proposed by Gorard (2013), described as a spiral with no clear beginning or end, with overlapping phases occurring simultaneously and could iterate almost endlessly. The cycle describes seven phases in the research cycle, progressing from evidence synthesis (phase 1), development of an idea or artefact (phase 2), feasibility studies (phase 3), prototyping and trialling (phase 4), field studies and instrument design (phase 5), rigorous testing (phase 6) and the dissemination, impact and monitoring of the intervention (phase 7). The cycle is based on a number of sources, including the genesis of a design study (Gorard & Taylor, 2004), research and design (Gorard, 2013) and the UK Medical Research Council model for undertaking complex medical interventions (MRC, 2000).

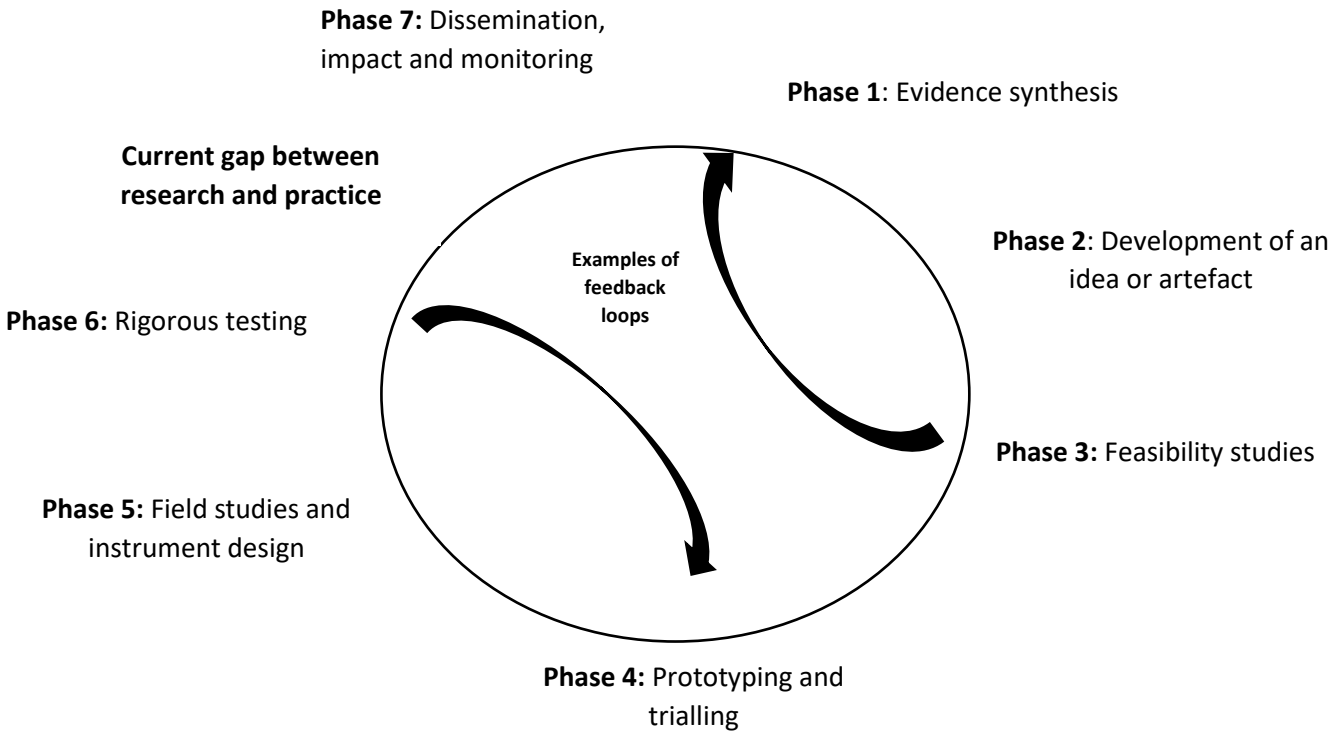


Figure 1: Adapted model of the outline of the full cycle of social science research and development (Gorard, 2013)

The research cycle demonstrates the phases involved in developing research from an initial evidence synthesis through to rigorous trials involving large scale RCTs and the dissemination of evidence through organisations such as the EEF and WWC. The advantage of disseminating research through these types of organisations allows research to reach the practitioners, increasing engagement and impact of the research findings. As practitioners do not have access to research journals, organisations such as the EEF and WWC provide a centralised method for disseminating research and bridging the gap between theory and practice. In each of the models discussed for phases of research, the concerns raised by Ginsburg & Smith (2016) relate to the external validity of RCTs in educational settings, are focused on phase 6 (rigorous trials) and phase 7 (dissemination, impact and monitoring) in the dissemination to classroom practitioners. At present, a disconnect appears to be evident between phases 6 and 7, with a lack of a mechanism to monitor the impact in terms of school usage, intervention effect size in participating schools and impact over a period of time.

As a practitioner does not often have access to academic journal articles this research cycle could be slightly contradictory, by implying that research is done by academics and disseminated to practitioners, thus creating a gap between academia and practice. If the cycle could involve the co-production of research with academics and practitioners working alongside each other, we could reduce increase levels of research understanding in schools and allow less friction in the dissemination of evidence.

At this point in the thesis, a limitation must be acknowledged to set out a justification for the primary research studies in this research focusing only on the testing stage of the research cycle and not addressing the dissemination phase. It was not possible within the timescale and limitations of resources for a PhD thesis to conduct primary research to test the complete theory of the model across all seven phases. As the cycle starts with the generation of evidence through the replication of small-scale trials, a pragmatic decision was made to start with the evidence generation and the use of protocols and small-scale RCTs linked to a prospective cumulative meta-analysis for the proposed testing phase of the research cycle. In future studies, the use of this methodology to address the issues raised around the dissemination phase will be explored.

2.3.3 The precursors of the What Works Centres – Cochrane and Campbell Collaborations

The UK Cochrane Centre was established in 1992 with the purpose of creating an international network for maintaining and preparing systematic reviews of the effects of interventions used in healthcare. The movement towards better informed policy and practice had not been new in 1992, however a growing number of researchers within the evidence movement had been campaigning for the knowledge base to be organised into a reliable and usable format. In the seminal text 'Effectiveness and Efficiency' (Cochrane, 1972), Archie Cochrane urged practitioners to practice evidence-based medicine and through the development of 'meta-analyses' it became possible to critically apprise and synthesise research findings in a systematic manner. The establishment of the Cochrane created an independent network of researchers, professionals, patients and carers working together to produce unbiased credible and accessible health information. The reviews published by the Cochrane Collaboration summarise the best available evidence to allow practitioners to inform their decisions based on robust evidence.

The National Institute for Health and Care Excellence (NICE) became established in 1999 to use the evidence from the systematic reviews in health to reduce the variation in the availability and quality of NHS treatments in the UK. NICE as an example of how systematic reviews of RCTs give rise to recommendations which are communicated to practitioners through guidelines and also used by service commissioners.

In the late 1990s, Iain Chalmers, a cofounder of the Cochrane Collaboration created a group to explore how evidence from the social sciences could be synthesised, similar to the evidence base from healthcare. Chalmers arranged an initial meeting in 1999 to explore the idea further, leading to the inaugural meeting of the Campbell Collaboration with 85 people from 13 countries in 2000 (Littell & White, 2018). The younger sibling to the Cochrane Collaboration, the organisation focused on

promoting the use of rigorous research synthesis methods through publishing guidelines and standards for systematic reviews in the social sciences. The mission of the Campbell Collaboration was simple, to promote positive social and economic change through the production and use of systematic reviews and other evidence synthesis for evidence-based policy and practice (Campbell Collaboration, 2019). Within the Campbell Collaboration, a number of specific groups have been formed to focus on areas within the social sciences, such as Crime and Justice CG, Education CG and the Social Welfare CG. The Education Campbell Group produces reviews on diverse issues in early childhood, elementary, secondary, and postsecondary education; topics include academic programs, teacher qualifications, testing, and a wide variety of school-based interventions (Littel & White, 2018).

The Campbell Collaboration has also evolved to create plain language summaries of the Campbell reviews designed for consumers through the development of the new Knowledge Transfer and Implementation (KTI) Campbell Group. The importance of the development of the Cochrane and Campbell Collaborations paved the way for the movement towards the What Works Centres for educational research, so these should be considered the precursors for the global movement towards evidence-based education.

2.3.4 What works centres – Global movement towards evidence-based education

In education, there has been a worldwide increase in the use of evidence in education to inform policy and practice. In the USA, bodies such as the Institute Educational Sciences' (IES) What Works Clearinghouse (WWC), the Best Evidence Encyclopaedia (BEE) and the Department General Administrative Regulations (EDGAR) have been established to inform education decision making. The IES is the statistics, research and evaluation arm of the U.S Department of Education. The purpose of the IES is to provide scientific evidence in an accessible format to educators, policymakers, parents, researchers and the general public. The evidence is independent and non-partisan, with the purpose of informing policy and practice based on scientific evidence. The WCC is a free resource funded by the by the IES, providing a connection between research and practice to improve education.

In the UK, the EPPI Centre and Education Endowment Foundation (EEF) have been set up in an effort to accelerate the movement of education towards evidence-based practice. The EPPI centre is based in the UCL Institute of Education at University College London, specialising in developing methods for systematic reviewing and synthesis of evidence, as well as developing methods for the study of the use of research. The EEF is an independent grant-making charity dedicated to breaking the link between family income and educational achievement. The charity was founded by the Sutton Trust in partnership with Impetus Trust, with a grant of £125 million from the Department of Education. The charity funds and independently evaluates educational interventions, investing in

evidence-based projects designed at tackling the attainment gap between disadvantaged students and their peers. The creation of the Sutton Trust–EEF Teaching and Learning Toolkit supports teachers, school leaders and policy makers make informed decisions about practice based on robust evidence. The development of the Teaching and Learning Toolkit (Higgins et al., 2015) is specifically designed to evaluate interventions, through large scale RCTs and meta-analyses of existing evidence. The toolkit provides a security rating and cost per pupil for specific interventions, allowing teachers the opportunity to rank interventions based on cost, impact and security of findings. The toolkit is different to previous efforts to by the Campbell Collaboration, as the toolkit is designed to communicate the findings of systematic reviews in a way that is understandable and accessible. In Scandinavia, the Norwegian Knowledge Centre and the Danish Clearinghouse for Educational Research have been established with the purpose of using evidence to inform practice. In Australasia, The Australian Teaching and Learning Toolkit and New Zealand Best Evidence Synthesis provide a similar purpose, ensuring the evidence available to policy makers, school teachers and leaders has been evaluated and met rigorous standards of review. It is important to note as the global evidence movement in education as grown over the last 20 years, so has the number of independent versions of evidence bases across the globe.

A common theme in these organisations over the last two decades has been the focus on using Randomised Controlled Trials (RCTs) as the best way to evaluate educational interventions at scale often labelling these the 'gold standard' for evaluations (Ginsburg & Smith, 2016). The increase in the use of systematic reviews and meta-analyses in education has seen a more rigorous approach to evidence collection, combined with the increase in experimental studies to develop an evidence base for practitioners. Yet, as stated by Rothstein et al. (2005), confidence in meta-analytic and systematic reviews are dependent upon the extent to which these findings are accurate and free from bias.

Two recent studies (Cheung & Slavin, 2016; Ginsburg & Smith, 2016) have raised methodological concerns and potential threats to the usefulness of a number of RCTs and meta-analyses. The authors of the paper do not have issues with the logic around the potential of RCTs in the field of education, as they recognise that a well implemented RCT is a good tool for establishing unbiased estimates of causal effects. As RCTs have strong internal validity, we do have potential issues related to the external validity, as these are often carried out for a particular sample, in a particular setting and a specific period of time. Therefore, the external validity of RCTs can be a potential issue. The issue of external validity will be addressed in further detail later in this chapter. However, it is important to understand the process of generating evidence in the social sciences and the potential gap highlighted by Hargreaves (1996) between researchers and practitioners.

2.3.5 How is evidence generated by the EEF / WWC and the important questions that need to be asked?

The purpose of the WWC and the EEF teaching and learning toolkit is to provide evidence for schools by finding out what works. Since the EEF was founded in 2011, 66 interventions have been independently evaluated with 90% of trials involved as efficacy or effectiveness studies. The charity aims to invest £200 million by 2026, raising the attainment of disadvantaged children by:

- Identifying and funding promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- Evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale;
- Encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective (Education Endowment Foundation, 2015)

The WWC provides a similar tool for teachers with the 'Find what works' search, listing educational interventions with an improvement index, effectiveness rating and extent of evidence. In the last decade, over 700 publications and more than 10,500 studies have been reviewed with the aim of informing researchers, educators and policy makers as they work toward improving education for students (Institute of Educational Sciences, 2015).

When considering phase 7 in the research cycle involving the dissemination, impact and monitoring of educational interventions the following questions identify the weakness in the current model of creating evidence. Once the initial RCTs and meta-analyses have been completed, evaluated and reported, the following questions are crucially important. These are:

- Which schools are implementing the intervention based on the research?
- How can we monitor the impact of the intervention in schools?
- How can we measure the impact of the intervention?
- How can we monitor the impact of the intervention over time?
- As a teacher, how can the intervention be implemented in schools and provide evidence of the impact on their pupils?

The fundamental problem of a lack of monitoring and measuring the impact of research is evident between schools and educational research when reflecting upon these questions. At present, the

publication of research reports is the final dissemination for many RCTs, yet as a school leader it is not clear how to implement the intervention or how to replicate an evaluation in their school. An EEF Octopus trial has investigated the use and engagement of teachers with evidence, the first trial used a range of high-quality resources to disseminate the research whereas the second used a light touch approach. The two trials found no evidence for improved literacy attainment at Key Stage 2 and the second trial found no increase in teachers use of research, suggesting light touch interventions are unlikely to be successful (Lord et al., 2017).

Often, interventions are released whole school wide with little thought of evaluating the impact robustly, as school teachers and leaders are not specialists in designing trials. The initial organisations funding the RCTs have no mechanism to record which schools implement interventions based on the initial research, or if this has a positive impact in these schools. Consequently, it is not possible to determine if the most effective interventions which have been trialled are currently being adopted by educational practitioners, leaders and policy makers.

A recent development in the United Kingdom has been the creation of research schools. The Research Schools Network is a partnership between the EEF and the Institute for Effective Education (IEE) to develop a network of schools to engage with research evidence to improve teaching (EEF, 2017). The purpose of the initiative is for the research schools to become focal points for evidence-based practice to be embedded into regions across the country, then through local partnerships to diffuse this into affiliate schools. At present, no evaluations have been conducted into the impact of these research schools on the diffusion of research evidence and how this is adopted by schools. The creation of the Research Schools Network is a positive step in attempting to close the gap between research and practice identified by Hargreaves (1996), however research into the dissemination research findings in education shows that this might not be sufficient in changing practice.

The next section discusses the research evidence on the strategies used to increase the use of research in practice, the theoretical models for evidence-based education and the conceptual models dissemination of research.

2.4 Research evidence on the strategies used for increasing the use of research

2.4.1 Models of evidence-based practice

Before discussing the potential barriers which may hinder the use of research in education, it is important to consider the different models associated with evidence based practice. Nutley, Walter, & Davies (2009) suggest three models of evidence-based practice; researcher-based practitioner

model, embedded research model and organisational excellence model. Even though these models were developed for social care by Walter et al. (2004), the authors suggest that these models are applicable to education.

Researcher based practitioner model

The underlying assumption in this model is that the responsibility for keeping up to date with the most relevant research and using this evidence to inform practice lies with the individual practitioner. This is a linear process, as the research is accessed, reviewed and applied to practice in largely instrumental ways. In this model, teachers are perceived to have a high level of professional independence in being able to critically appraise research to inform their own practice.

Embedded research model

The model moves the responsibility from the individual practitioner to the systems, processes and frameworks in the educational system. These include national educational policies, LEA guidance and Ofsted frameworks. In this model, research enters practice indirectly with no link between practitioners and researchers. The model relies on intermediaries to bridge the gap between research and practice, though this model can still be seen as linear and instrumental.

Organisational excellence model

This model assumes the responsibility for research informed practice lies with the organisational leadership, management and culture. Primarily, it is the organisations role to review external research findings and channel these into practice, as well as acting as a centre for local experimentation and evaluation. The model involves the co-creation of knowledge though organisations working in partnerships with Universities and other organisations such as the EEF in conducting primary research.

Nutley (2009) explains that the three models are helpful in shaping thinking about research strategies. Yet caution must be attributed to these models, as these are not mutually exclusive with some interventions bridging two models.

The three models each have advantages and limitations for increasing the use of research with practitioners, these will be briefly discussed. The researcher-based practitioner model places responsibility with the individual practitioner. In an ideal world whereby practitioners have the time to keep informed of research, having engaged practitioners evaluating their own research should allow research findings to be used as intended. However, in order for practitioners to be able to do this they require the knowledge to be able to critically appraise research. In practice, the implications and resources required to train teachers through continued professional development (CPD) to enable

them to effectively embed and evaluate the impact of interventions within their own practice might not be feasible or practical.

The second embedded research model moves the responsibility away from the individual practitioner to the systems and processes within the educational framework such as LEA's, policies or Ofsted. As the practitioner is not involved this model does not address the potential lack of knowledge or understanding within practitioners, relying on intermediaries to bridge the gap between research and practice. Consequently, as the model does not require intensive training, it has the potential to reach a large number of practitioners in a shorter timescale. Although, even if it reaches more practitioners, if they do not understand how to implement research we are unlikely to see any positive results in terms of translating positive effects from robust trials to practice.

The third model for organisational excellence assumes the responsibility lies with the leadership, management and culture. The model has the advantage of using schools as centres for local experimentation and evaluation, similar to the EEF research schools programme in the UK. The co-creation of knowledge with schools working with universities or organisations such as the EEF provides the opportunity for researchers and practitioners to work in a symbiotic relationship. As a model this allows researchers to upskill leaders in schools to disseminate knowledge to teachers, within the setting of their own schools. A limitation of this model is still finding a way to bridge the gap between understanding how to implement the research in an accessible and time efficient method due to the complex nature of schools.

If research leads in schools could be provided with an off-the-shelf protocol to follow as a step by step guide to implementing an intervention, with guidance on how to set this up as a robust evaluation, we might have a way to bridge the gap between theory and practice. The proposed model for the use of protocols and small-scale aggregated trials will explore this further in this chapter.

2.4.2 Conceptual framework for dissemination of research

The term dissemination is defined by Freemantle & Watt (1994) as the mechanisms and strategies by which specific groups become aware of, obtain and make use of information. In the research literature, many other terms are used, such as 'knowledge mobilization', 'knowledge transfer' and 'knowledge translation' (Levin, 2013; Mitton et al., 2007; Nelson & O'Beirne, 2014). The definition suggested by Freemantle and Watt (1994) introduces the assumption that dissemination involves the specific targeting of groups with research evidence, whilst also highlighting the requirement for these groups to engage and use this information. However, Levin (2013) suggests that the term 'knowledge

mobilization' defines the process more precisely as it indicates that the work requires specific effort, over time and working with others. He argues that the term 'mobilization' demonstrates that the process is not a one-way process, capturing the interactive, social and gradual relationship between researchers and practitioners.

The idea that dissemination involves the movement of knowledge to specific groups over a period of time, with the aim of the users accepting and adopting this knowledge into practice also links to the theory of diffusion. Rogers (2003) explains diffusion as a process by which an innovation is communicated through certain channels over time among members of a social system. The theory suggests that diffusion is a special type of communication, involving the communication of new information as a process whereby participants create and share information to each other in order to reach a mutual understanding. The word diffusion is often associated with the unplanned spread of new ideas or evidence, whilst the term dissemination infers a directed and managed flow of information. As previously mentioned, Levin (2013) prefers the term 'mobilization' as this term makes it clear that the process is not a one-way process. However, if diffusion is considered as a two-way communication then this point is not necessarily correct. Rogers (2003) explains that diffusion is a special type of communication, a two-way process of convergence involving a social change. This is defined as the process which alteration occurs in the structure and function of a social system. Consequently, diffusion and dissemination are often interchangeable as it is often not clear in actual practice if the spread of new ideas is planned or unplanned. Therefore, the word dissemination will be used conceptualise the idea of communicating evidence, as a two-way process of convergence involving a social change in this thesis.

This next part of this chapter provides current conceptual models on the process of dissemination. It is important to note at this point that numerous proposals have been previously proposed in an attempt to bridge the gap between research and practice (Lysenko et al., 2014), with examples including the work by Nutley, Percy-Smith and Solesbury (2003) and Landry, Amara, and Lamari (2001). Yet, these studies are limited by the lack of empirical evidence upon which they are based on (Hemsley-Brown & Sharp, 2004; Levin, 2013; Lysenko et al., 2014). It is important to try to understand the conceptual models for dissemination before briefly summarising the empirical evidence on the use of evidence by practitioners and the potential barriers encountered when disseminating research findings.

2.4.3 Conceptual models for dissemination

The conceptual framework for knowledge mobilisation (KMb) has been developed over a number of years in an attempt to conceptualise the main processes involved in KMb. Levin (2013) presents a conceptual model for the process of dissemination that identifies three overlapping and interacting domains. These consist of the production of research, the end use of research and the intermediary processes that link these. Figure 2 represents the conceptual framework, with triangles representing the functions and not necessarily structures.

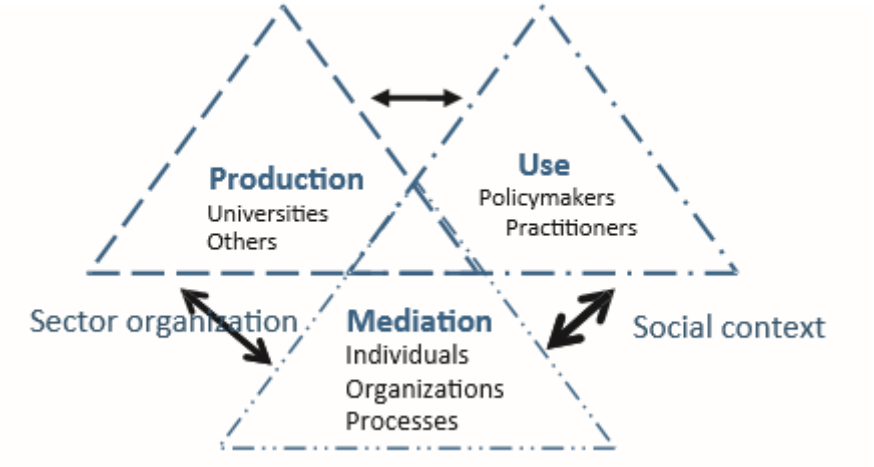


Figure 2: Conceptual framework for knowledge mobilisation (Levin, 2013)

Levin refers to these three as 'contexts' because particular individuals or organisations can fit into more than one of the contexts at any one time. For example, the production of knowledge is generally produced in Universities by researchers, who then act as intermediaries when promoting their own research to practitioners. Additionally, teachers involved in research can produce knowledge then act as intermediaries. The arrows between the three contexts show that these are all related to each other through a range of personal and organisational connections to varying degrees of intensity (Levin, 2013).

Importantly, Levin (2013) argues that while changing policy is generally difficult, it is in many respects less difficult than changing practice at an individual teacher level. In education, a small number of people can potentially change a policy but in terms of changing practice the education system is considerably harder to alter. It takes only a small number of people to make decisions and these are often people who do not have to embed the change into their own practice. It is the practitioners who have the responsibility of embedding the change in their practice so this will involve

a change in behaviour. Chinn & Brewer (1993) highlight the difficulties involved in changing beliefs and working practice of professionals even when presented with conclusive evidence that those ideas and actions are misplaced. Enabling this change may require many hours and different forms of CPD, with Adey et al., (2004) suggesting it takes approximately 35 hours of CPD to bring about meaningful change Therefore, the change must occur in large numbers of educators who are spread across many organisations.

Sharples (2013) presents a model based on an evidence chain as a connected ecosystem, adapted from Shepherd (2007). The complex chain involves research production, synthesis, distribution, transformation and implementation working in a connected ecosystem. Sharples points out that there is no use in creating world class research on education interventions, if the research is not accessible to practitioners then the research has been wasted. Hargreaves (1996) identifies a gap in evidence chain between researchers and practitioners, suggesting this is a fatal flaw in the model as it is the reserachers and not the practitioners who determine the research agenda. To some extent, Hargreaves criticisms directed at educational researchers are particularly harsh, as many researchers in education have previously worked in the school settings (Hammersley, 1997) and are required to have a diffferent skill set to practicing teachers in schools. It is not feasible to expect teachers to have the skills to conduct large scale RCTs and analyse complex data sets, however teachers should be sufficiently skilled to interpret findings and use their professional judgement about its use in their practice. This is the potential flaw in the system, as a lack of continual engagement between researchers and practitioners would allow a better understanding of each others perspecive. Creating the opporunity for researchers to work alongside pracitioners will draw the researcher closer to the needs of practitioners, whilst also providing profesional development to upskill the practitioner in therorectical perspectives.

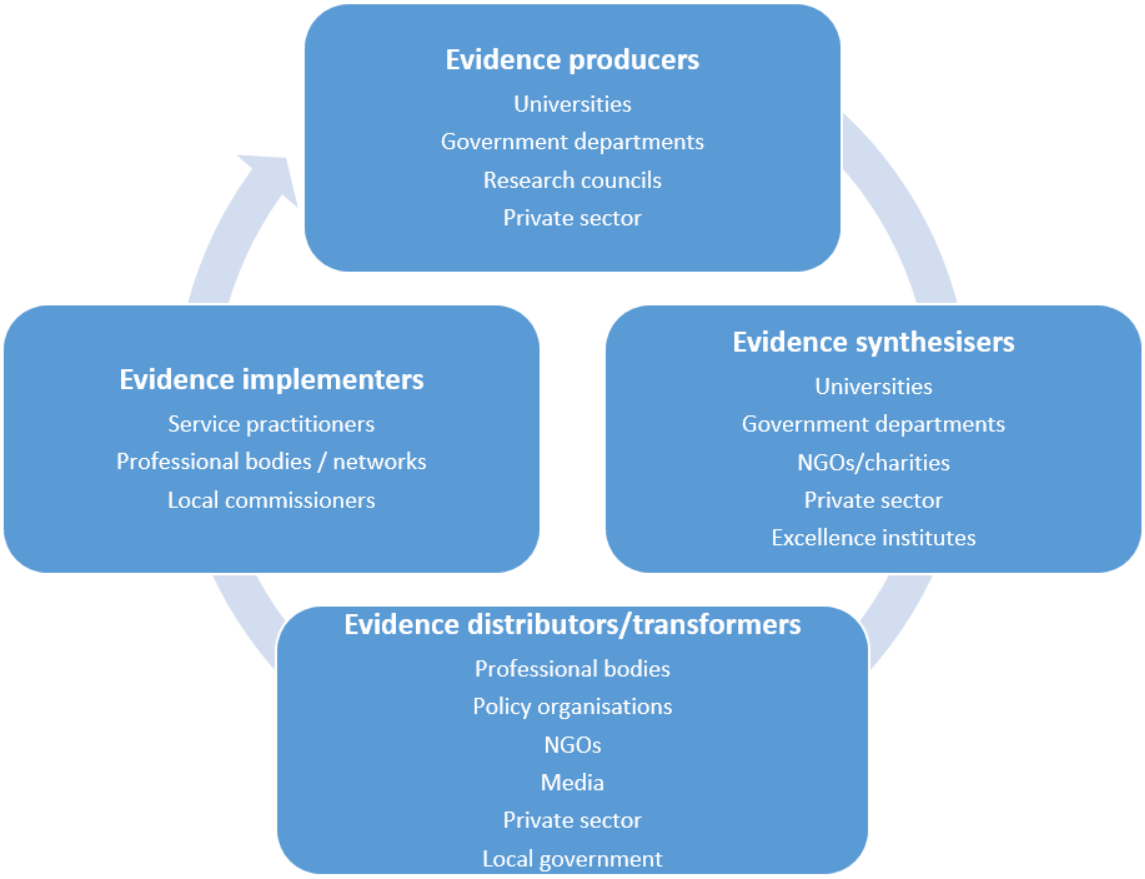


Figure 3: Outlines the elements of an evidence ecosystem, and this model will be compared with the model suggested previously by Levin (2013).

The models highlighted so far on the dissemination of research evidence (Levin, 2013; Sharples, 2013; Shepherd, 2007) demonstrate that it is important to consider these elements as a whole and not as individual components of generating evidence. A key theme emerging is the flexibility in the system, as a number of agents can act across multiple domains. Levin (2013) highlights this by stating that individuals and organsiations can fit into more than one context at anyone time. Additionally, a second theme is the importance of mediators between the evidence producers / synthesisers and evidence implementors. Relating back to the three models of evidence based practice presented by Nutley et al., (2009) as either researcher based practictioner, embedded research or organisational excellence models, it is important to try to understand the role of mediators. For example, a teacher may be skilled at interpreting evidence and attempt to implement a new strategy based on evidence. If the organsiation that they are part of is not supportive, in terms of time to read such evidence or does not support innovation as they have whole school or academy chain procedures, then they will not be

allowed to change practice based on research evidence. In order to effect a change, a mediator may be required to change organsiational practice.

In an attempt to understand how research evidence can be disseminated effectively to practitioners and decision makers in their organsiations, it is useful to consider the work of Rogers (1983). Rogers synthesized research from obver 508 diffusion studies across a number of research fields in 1962 and produced a theory of the adoption of innovations amoung individuals and organisations. The diffusion of innovation theory suggested by Rogers (1983) can be applied in the context of education, particularly in the knowledge generation involved in determining what works in relation to educational interventions. The original model defines innovation as an idea, practice, or object perceived as new by the individual or another unit of adoption. It is the characteristics of the innovation, as perceived by the social system, which determine the rate of adoption within this system

Rogers (1983) claims that the adoption of new ideas is a process of social change, identifying two key roles. Firstly, *change agents* are people with professional training who are involved in bringing about a wider adoption of the new idea. Secondly, *opinion leaders* are influential people within a field who are influential in their social networks. If a new idea or research evidence is to be adopted into practice, the effective dissemination can be achieved by identifying opinion leaders and concentrating change agents contact on them. The majority of individuals evaluate an innovation, such as an educational intervention, not on the basis of scientific research evidence, but through the subjective evaluations of near-peers who have adopted the innovation.

The importance of the diffusion of innovation theory proposed by Rogers (1983) will be explored further in chapter 8 of this thesis, lining to the micro - , meso – and macro-levels within the education system. Furthermore, the theoretical and practical limitations of an aggregated trials model linked to a prospective cumulative meta-analysis as a method for disseminating and evaluating school based impacts will be discussed.

To recap, the process of disseminating research involves a complex ecosystem involving producers, synthesisers, transformers and implementors. The individuals and organsisations are not context specific, as these can fit into more than one context at anyone time. As dissemination is not a passive process, the social nature requires identifying change agents and opinion leaders to support the impelmentation of the innovation or intervention. In the context of educational effectivess studies, a fundamental requirement in the design of the study should be the processes involved in disseminating the findings and how these can be made accessible to practitioners and policy makers. In educational research, a number of academics have raised the issue of a disconnect between conducting research and embedding the findings into practice and policy (Lysenko et al., 2014), it is

therefore necessary to review the empirical evidence base related to the use of research in education. A number of theorectical models for bridging this gap have been suggested, yet these are not based on empirical evdience (Hemsley-Brown & Sharp, 2003; Levin et al., 2011).

2.5 Empirical evidence for research use in education

The empirical evidence for how educators value, perceive and use research to inform practice and policy is limited. The following section summarises the empirical studies which have focused on the use of research by teachers and the barriers that can affect this use.

2.5.1 Evidence on the use of research by school practitioners

A number of factors relating to the use of research within the teaching profession have been identified in the literature. The first perspective to be considered is from an information literacy perspective. Williams & Coles (2007) examine the use of research by school practitioners with a focus on their information literacy, the ability of teachers to find, evaluate and use research findings. The authors study builds on the earlier research on the two associated factors which are known to influence information-seeking behaviour: access and attitudes (Wilson & Walsh, 1996). The research included a survey of 3000 school teachers and 500 head teachers in Scotland, England and Wales, with interviews conducted with 28 practitioners.

Prior to discussing the research findings, it is important to acknowledge the limitations of this research. Attrition rates in the survey responses were incredibly high, with only 10.9% of teachers and 15.6% of head teachers completing the survey from the sample. Therefore, concerns regarding bias are high due to the nature of teachers who responded. Are these the types of teachers who would usually engage with research as they had the time to engage with this study?

The study by Wilson & Walsh (1996) found that teachers lack of research literacy may be a factor in limiting the use of research information and that this is exacerbated by a lack of time and access to research. The findings reported between 60 – 80% of respondents from UK schools used research from ‘never’ to ‘occasionally’. The authors suggest three strategies to improve research use through; developing an information culture and ethos in schools, increasing local information dissemination strategies and a greater development of teacher information literacy skills. As

previously mentioned, the high attrition in the study does raise concerns relating to the representative nature of the sample and the generalisations that can be made from this single study.

Cousins & Walker (2000) conducted a small-scale exploratory study to investigate the variables which best predicted teacher's attitudes toward research use in schools. The study surveyed 310 educators in a sample of schools in Canada as part of a wider research project. The variables were selected as potential predictors of educators' self-reported views about applied research utility and relevance: these included personal teaching efficacy, prior participation in research, prior research coursework, gender, experience, organisational learning capacity and the type of school (elementary or secondary). Two hundred and ninety-seven respondents completed the survey and a stepwise multiple regression model was used by the authors to determine predictor variables for research use in the teaching profession. It should be noted that the sample size is relatively small and the research was conducted as an exploratory study. Yet the findings suggest that prior participation in research and sense of personal teaching efficacy both consistently explained variance across all five composite measures of attitudes toward research use (Cousins & Walker, 2000). This is supported by previous studies that suggest that teachers involved in research find their participation rewarding in terms of developing professionally as teachers and improving their understanding of the use and purpose of research (Cousins & Leithwood, 1993; Noffke, 1992). It is important to note that Cousins & Walker (2000) found that school teachers in their sample rarely read research papers.

Interestingly, the authors found that the number of years of teaching experience resulted in having negative opinions about their own research abilities and other teacher's participation in research. Possible reasons for are discussed in the paper and these include teachers reaching a career stage whereby they "disinvest" in school work to avoid additional teaching tasks or extra work commitments. Other reasons include school culture and organisational structures stifling teachers or as they gain experience, they become entrenched in routines and cultural norms so become resistant to change.

Hemsley-Brown & Sharp (2004) conducted a systematic review into the use of research to improve professional practice, exploring how teachers use research, which features of research encourage teachers to use the findings in their practice, whether healthcare practitioners make greater use of their findings and approaches to dissemination. The studies included in the review were published between 1988 and 2001, with a total of 21 empirical studies included in the review. However, the studies included have the potential of high risks of bias due to their research design and the generalisations which are implied in the review. For example, the author states "the most pertinent evidence for this review is provided by Zeuli (1994). This empirical study specifically

concentrated on the way teachers use research findings. The study was carried out in Michigan in the United States with a convenience sample of thirteen primary, middle and secondary school teachers, and aimed to find out how teachers read and respond to educational research" (Hemsley-Brown & Sharp, 2004). The sampling method and size (13 teachers) raises a number of concerns, as well as the author referring to this research as the most pertinent evidence for this review provides little confidence in the robustness of this evidence.

Lysenko et al. (2014) conducted an investigation into the predictors of school practitioner's use of educational research, as part of a research strand evaluating a governmental initiative in Quebec, Canada. The sample included 2,734 questionnaires returned anonymously, with a participation rate of 58.7%. Overall, the respondents positioned their use of research at the low end of the scale, between "never" and "once or twice" during the last year. This finding support the earlier studies in the low use of research by school practitioners (Cousins & Walker, 2000; Williams & Coles, 2007). Interestingly, Lysenko et al. (2014) used Roger's model of innovation diffusion (1995) to group the range of factors that are likely to affect a practitioners use of research in their practice based on empirical studies. These four factors; practitioners opinions about research, expertise, organisational factors and attitudes towards awareness activities consistently and modestly explained the frequency of the use of research by teachers. Of these factors, the attitudes of teachers toward research was the largest predictor of use.

Nelson & O'Beime (2014) conducted a rapid evidence review for the National Foundation for Educational Research (NFER) into the effective approaches to school and teacher engagement with research evidence. The systematic review included UK literature on research use in schools since 2010, yet the authors acknowledge the incredibly scant evidence base as no studies involved randomised controlled trials. The authors endeavoured to select the most methodologically sound studies for analysis, however these were mainly based upon observational studies or small-scale qualitative research designs. A key finding in the research is that there is a need for better evidence of the impact of different approaches to disseminating research findings to transform practitioner knowledge. "The evidence base should be established before too much more resource is invested in developing a flow of evidence for practice" (Nelson & O'Beime, 2014).

Recent research by the EEF through their Research Use in Schools funding stream has evaluated five projects to tackle the evidence gap in disseminating research findings into practice. The first study involved evidence-informed CPD in Rochdale, involving 10 primary schools and approximately 280 pupils. The objective of this study was to "explore whether, and to what extent, research communication and engagement strategies had the potential to improve teachers' use of,

and attitudes towards, academic research to support pupils' progress" (Speight et al., 2016). The pilot study found positive changes in teacher's attitudes towards research but no evidence that the teachers were more likely to use research to inform their teaching.

The second study involved a pilot project run by the Ashford Teaching Alliance, testing whether a 'research champion' working across five schools to improve the awareness and use of evidence in the classroom is a feasible model. The study found no evidence that teachers attitudes towards research or their use of evidence changed during the intervention. It was also noted that attendance and engagement was often low due to time pressures faced by teachers (Griggs et al., 2016).

The third study (Research Learning Communities) was an efficacy trial of a project developed by the Institute of Education to examine whether evidence champions are effective at promoting research use in their school when supported by a research community of peers from local schools and an academic facilitator. The project found no evidence that Research Learning Communities improves reading outcomes for children at Key Stage 2. The project did find some evidence for a positive impact on teachers' disposition towards research, yet this could be influenced by other factors in the selection process of the teachers involved (Rose et al., 2017).

The fourth study The Literacy Octopus – Communicating and Engaging with Research: a large multi-arm randomised controlled trial investigating a range of different methods of communicating research to schools and engaging them in research evidence. The study found light touch dissemination strategies are unlikely to have an impact on the use of research by teachers (Lord et al., 2017).

The fifth study is The RISE Project – Evidence-informed school improvement: an efficacy trial of a project led by Huntington School that aims to test whether a research-based school improvement model makes a significant difference to classroom practice and student outcomes. The study found that there was no evidence that the programme had a positive impact on the outcomes of pupils who were eligible for free school meals. More importantly, the schools' adoption of the research-informed school improvement model was highly variable and influenced by schools' context and relationships, and the stability of the Research Lead role. In the study, the teacher in the role of Research Lead changed in 40% of the schools during the project (Wiggins et al., 2019).

It is clear from the limited research evidence that the use of research in the teaching profession is low. For this reason, it is important to identify the potential barriers to the use of research by school practitioners.

2.5.2 Barriers to the use of research by school practitioners

The barriers to increasing the use of research in educational institutions has been documented in a number of studies (Levin et al., 2011; Lysenko et al., 2014; Mitton et al., 2007; Nutley et al., 2009). In order to explain these potential barriers, it is useful to divide these into barriers related to research itself and the knowledge and skills of practitioners.

Levin et al. (2011) identifies the barriers associated to the research itself to include inaccessibility, inconsistent results and unclear implications of the research for practice. Inaccessibility can be considered on two levels, firstly through the process of knowledge production of research. In research, the peer review process creates a barrier for practitioners as access is often restricted to academic journals. Secondly, inaccessibility can relate to the ability of the reader to understand the research as often academic journal articles are written for the purpose of the research community and not practitioners. A number of other studies highlight the inaccessibility and insufficient contextualisation of research (Bannister, 2011; Proctor, 2015; Sharples, 2013). An example of good practice in avoiding these issues is the publication of all the randomised controlled trails commissioned by the EEF on their website and the inclusion of a two-page executive summary which is accessible to teachers. However, this is often not the case with peer reviewed journal articles.

Another potential factor which is a problem rather than a barrier, it that research is used but not in the way it is intended. If a practitioner adapts the intervention to fit within their own context, then the intervention may not be effective. In this instance, it is not the understanding of the academic paper but the lack of the understanding in how to replicate intervention as it was intended to be used.

Inconsistency in the results and poor quality research often leaves schools and teachers to assess the quality and findings of single studies of varying quality, making it challenging to understand which research to implement in practice (Gough, 2013). Furthermore, Levin et al. (2011) highlights the barriers are created by a lack of knowledge and skills of practitioners, these include practitioners may not know where to find the most current research, they lack the skills to assess the quality of the research and practitioners are unable to interpret the findings if they are expressed in terms of effect sizes and significance levels.

At the organisational level, factors such as the school setting and context, culture and professional ethos (Cousins & Walker, 2000; Hultman & Horberg, 1998) can affect a schools readiness and capacity to support individual teachers to engage with research. If schools do not provide time and opportunities for teachers to read and engage with research, it is less likely that research will find its way into everyday practice. Furthermore, schools are embedded into a larger system which has a

number of external factors such as government policy, external agency pressures such as Ofsted, funding and teacher recruitment issues in schools.

2.5.3 Limitations of the current evidence

It is ironic that the debate over the use of research is itself not informed by reliable evidence. The majority of evidence as described by Nelson & O'Beime (2014) as an incredibly scant evidence base, with little evidence using randomised or quasi-experimental research designs. The majority of the research relating to the use of research to inform practice involves surveys and interviews, relying on practitioners to offer explanations that are ex post facto. In these instances, Levin (2013) argues ideas may influence their beliefs and practices gradually over time in such a way that the original connections are no longer visible to the participants being interviewed.

The current work of the Education Endowment Foundation through the research use in school studies will hopefully provide a basis for further trials into the effective dissemination of research findings into schools. More work is needed in the field of academic research to create an evidence base to allow research to inform practice, as this is the purpose of educational research evaluations of interventions. Finally, it is important to acknowledge that the effective dissemination of the research findings to bring about a change in practice should not be the end of the research cycle. Crucially, the key question on monitoring the impact of these changes in school settings and replicating findings is just as important as disseminating the findings in practice.

2.6 Aggregated trials in education

In the context of this thesis, an aggregated trial is defined as the aggregation of separate small-scale randomised trials following a pre-defined protocol over sequential time period. In educational research, traditional randomised controlled trials are conducted at one specific time point. At present, no evidence exists on the use of aggregated trials as defined above, yet randomised trials have been conducted at pupil level within schools and the results aggregated from the participating schools in the sample. These trials are not described as aggregated trials, differing from the proposed model as the data is collected at one point in time. An early example of such a trial is the Every Child Counts evaluation (Torgerson et al. 2011) involving 44 schools in the UK. The intervention involved a 12-week one-to-one mathematics programme delivered daily, with 12 pupils randomly assigned to either the autumn, spring or summer allocation of the intervention. The pupils assigned the intervention in the

autumn term received the intervention, with the post test results of the spring and summer acting as a control group and allowing the effectiveness of the intervention to be assessed.

It is not possible to identify how many randomised controlled trials have used individual randomisation and aggregated the results, as no systematic reviews into the design of randomised trials has been conducted. The evidence for the use of aggregated trials (RCTs) in education is limited, with only two recent evaluations conducted through the EEF using this approach. However, these were conducted at only one time point so could be classed as a randomised control trial with individual randomisation at school level. Accelerated Reader (AR) is a widely used web-based reading intervention used in over 2000 UK schools. However, it was not clear that the implementation of the intervention at such a large scale could be justified solely on the basis of the existing evidence of effectiveness (Siddiqui et al., 2015). Four individual schools proposed to conduct the intervention and evaluate the impact of AR in their schools. Following discussions with the EEF and independent evaluators, the decision was made to allow all four schools to conduct individual trials with the results aggregated for the final analysis.

The trial was school led with advice from evaluators, with a randomisation completed by the schools following advice, and checked after the randomisation to ensure the two arms of the trial were initially balanced in terms of KS2 scores in English. A waiting list design did not deprive pupils of the intervention, with the treatment group receiving AR first and the pupils assigned to the control continuing with normal lessons. The control received the AR intervention after 20 weeks when the trial was complete.

The trial demonstrated a + 0.24 effect size for the AR intervention compared to the control group using a school led aggregated trials model. However, more importantly the study demonstrated the feasibility of running school led experimental studies and aggregating the results. Siddiqui et al. (2015) explains the advantages of schools running their own trials, including lower cost, easier permission to innovate, and less drop out. The training involved with the schools also helps teachers consume other evidence critically.

The second trial in education involving the aggregated trial model evaluated a phonics intervention. The trial was conducted by the same authors of the first example (Gorard, Siddiqui, & See 2015). These two evaluated EEF trials using aggregated school led trials have demonstrated the capability for schools to run their own trials with appropriate guidance, and other bodies can aggregate the results to create findings with sufficient scale to be taken seriously.

The use of aggregated trials in education provides an opportunity for small scale aggregated trials to be conducted in school settings. Siddiqui, Gorard, & See (2015) have demonstrated the feasibility of schools to manage these types of trials, with the results of the mini-studies aggregating together to

increase the sample size, providing a more robust evaluation of the effectiveness of the interventions. However, the aggregation occurred at one point in time and not over a sequential time period, as well as involving a small sample of schools. Yet, the importance of this study should not be under-estimated as it provides empirical evidence for the basis of this thesis, that small scale aggregated trials managed by schools under the supervision of academic researchers can be conducted in school settings. It should then be possible to aggregate these studies over a sequential time period, using a procedure used in the healthcare called a cumulative meta-analysis.

2.7 The use of cumulative meta-analyses in healthcare

Cumulative meta-analyses is a method that could have a profound impact on the practice of education and on the settings of educational policies in the next decade. Lau, Schmid, & Chalmers (1995) defined a cumulative meta-analysis as a product of performing a new meta-analysis every time a new trial is added to a series of trials. Therefore, at each 'wave' of the database (each time a study is added), a separate meta-analysis is conducted (Higgins, Whitehead, & Simmonds, 2011). The accumulation may proceed using a number of covariates, such as sample size, year of publication, size of the difference between the treatment and control, social economic status, gender or any other criterion. The studies can be pooled in ascending or descending order.

In medicine, the earliest example of cumulative meta-analyses allowed researchers to show the use of intra-venous streptokinase for acute infarction to be a lifesaving therapy 25 years before its approval by the Food and Drug Administration (Lau et al., 1992).

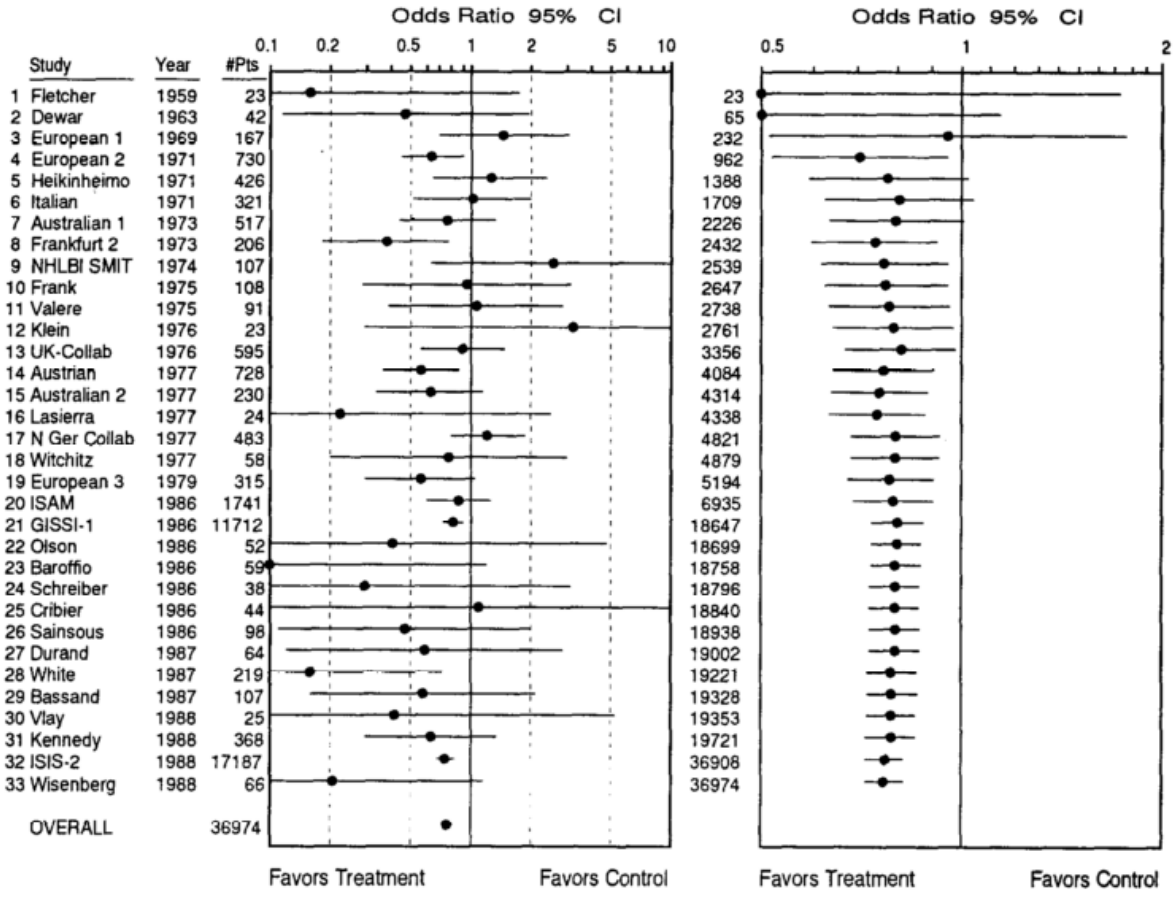


Figure 4: Comparison of a typical meta-analysis (right) of intravenous streptokinase for acute myocardial infarction with overall mortality as the end point. The Mantel-Haenszel method of pooling was used. Odds ratios and their respective 95% confidence intervals of individual studies and the overall estimate are plotted on the right graph p.250 (Lau, J., et al., 1992).

Figure 4 shows that the probability that the treatment was effective was at least 0.975 (97.5%) by 1973, with a cumulative randomised sample size of 2432 patients. At this time Lau (1992) states there was a 50% chance of that the treatment reduced mortality by at least 20%. By 1977, 4084 patients had been randomised with a placebo control or treatment, the probability that treatment reduced mortality by about 10% was 97.5% and the probability that the treatment was effective was more than 99.5%. Using a frequentist framework (rather than Bayesian) this resulted in significance levels achieving 5% in 1973 and 1% in 1977. Stampfer (1982) reported these findings in a meta-analysis yet randomised trials with treatment and placebo continued until 1988, raising major concerns regarding the use of placebo if effectiveness had already been established.

The addition of two very large trials, GISS-1 (11,712 participants)[1] and ISIS-2 (17,187 participants)[2] had no effect on determining whether or not the treatment was effective. The only impact of these two large scale RCTs was to increase the precision of the estimate of the treatment effect. Figure 5 demonstrates the concept of the cumulating data from trials, however the graphical presentation of the data using different scales makes interpretation difficult.

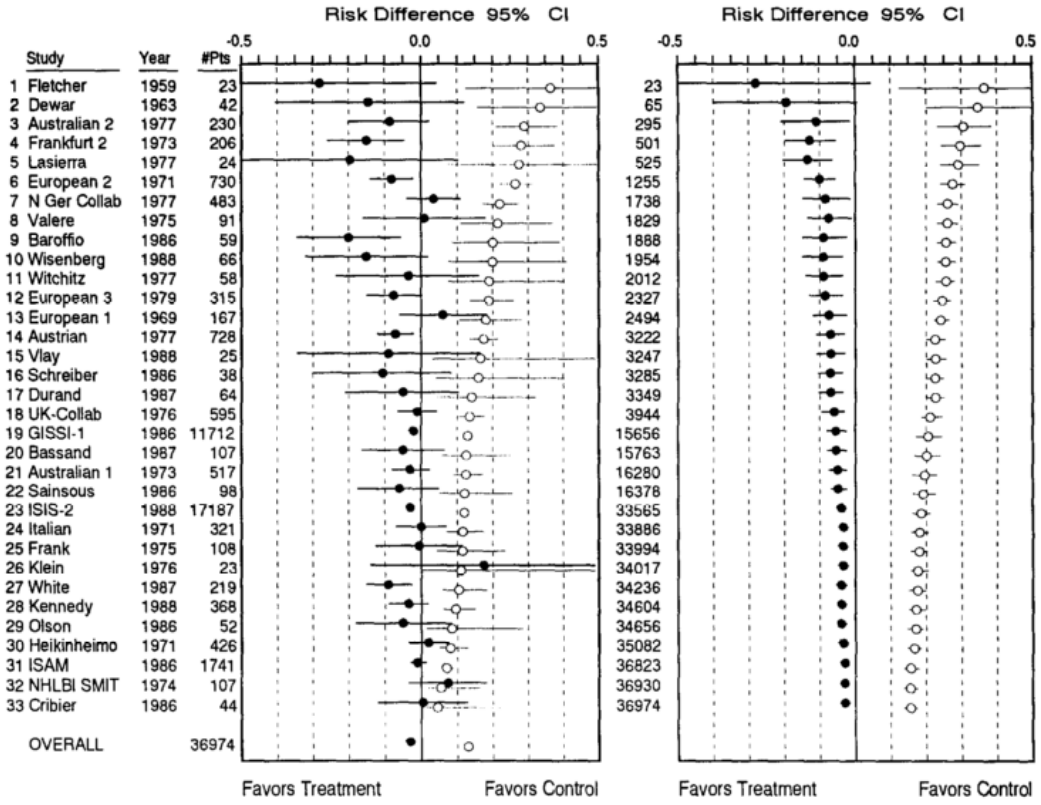


Figure 5: Reduction of overall mortality by intravenous streptokinase for acute myocardial infarction. Studies were pooled by the DerSimonian & Laird random effects risk difference method and arranged by decreasing control rate. Risk differences plotted as filled circles, control rates as open circles along with 95% confidence intervals. Individually plotted on right and cumulatively on the left graph using random effects model.

Figure 5 applies the same data from the previous trial using intravenous streptokinase studies but orders these by decreasing control rates. Pooling was completed by a random effects difference method (DerSimonian & Laird, 1986) and demonstrates the control rates of the individual studies are heterogeneous with a range from 5% to 36%. The advantage of using the cumulative meta-analysis are evident, with studies with high control rates being more likely to demonstrate higher estimates of

[1] (GISSI, 1986) Effectiveness of intravenous thrombolytic treatment in acute myocardial infarctions
[2] (ISIS2, 1988) Randomised trial of intravenous streptokinase, oral aspirin, both, or neither collaborative group trial

treatment efficacy. Conversely, low control rate studies are more likely to demonstrate no treatment effects or trends favouring the control (Lau et al., 1995).

Many recent examples exist in medical research (Cook, 2014; Yan et al., 2014; Zhang et al., 2014) on the benefits of cumulative meta-analysis but how can this be applied to educational research?

At present, no examples exist in the literature for the use of pre-planned prospective meta-analyses linked to small scale aggregated RCTs with pre-defined protocols. The following section outlines an introductory model for how this can be conducted in educational research.

2.8 An introductory model for evidence-based practice in schools using aggregated trials linked to a prospective meta-analysis

In education, RCTs provide a robust design for inferring causation if they are implemented well (Ginsburg & Smith, 2016). A well designed RCT has very high internal validity, however as previously stated these as are often carried out for a particular sample, at a specific time, particular place and in a specific setting. However, these studies are inherently expensive and often time consuming, with the EEF spending approximately £500,000 per trial. In a study by Lortie-Forgues & Inglis (2019), the authors found that the average effect size from commissioned RCTS with good methodological rigor was 0.06 standard deviations. When commissioning large scale RCTs, the EEF will test interventions as feasibility or efficacy studies, enabling promising studies with positive effect sizes to progress into the robust testing phase involving independent evaluators using RCT research designs. Yet, when scaled the effect sizes are uninformative and unable to replicate the previous studies.

If we consider the research cycle outlined by Gorard (2013) in Figure 6, Lortie-Forgues & Inglis (2019) suggest three plausible explanations on why effect sizes do not replicate between smaller efficacy studies and large scale RCTs. These include the possibility that the initial research evidence is flawed, a lack of translation in the initial insights during the implementation in the larger trial and potential underpowering of trials due to background noise. Each of these will be discussed in further detail in the final chapter, however it is important to acknowledge the problem we currently have in developing the evidence base on which to fund robust evaluations.

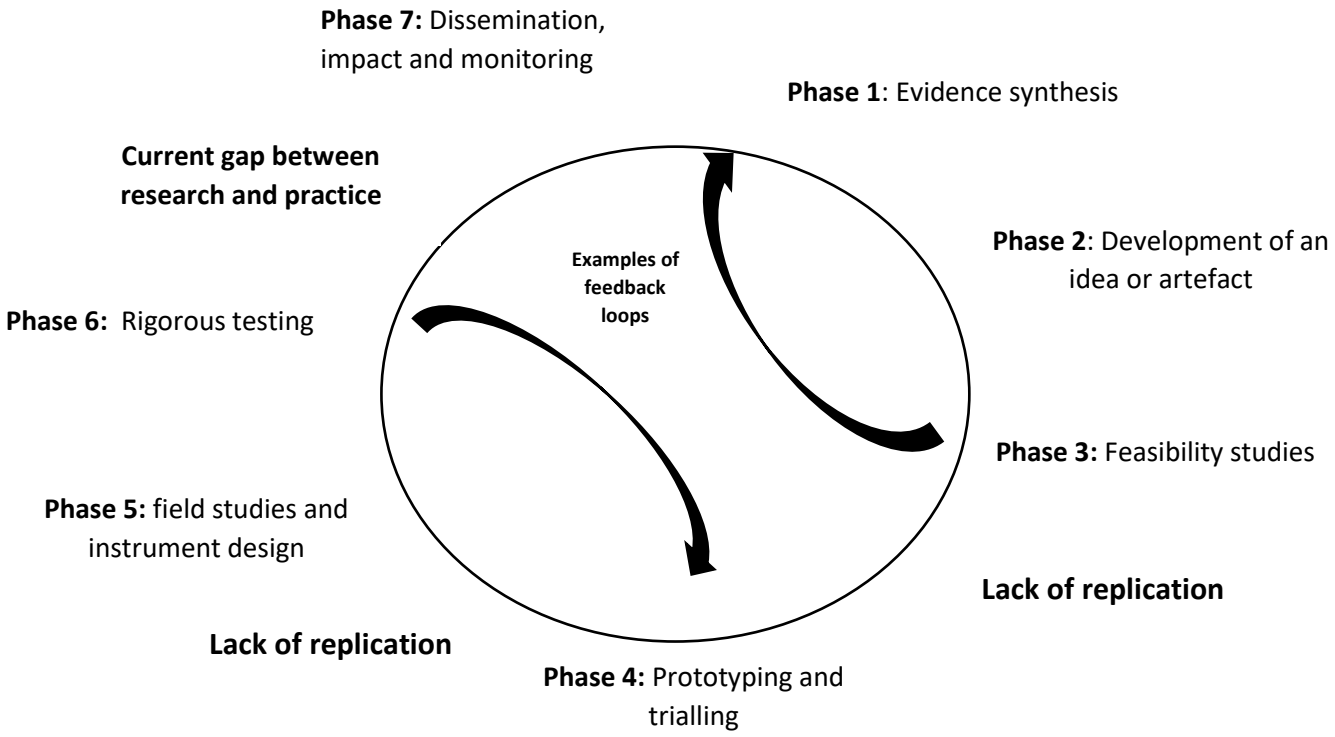


Figure 6: Adapted model of the outline of the full cycle of social science research and development (Gorard, 2013)

The purpose of the thesis is to propose a simplified version of the research cycle to combine the phases 3 – 5 into a testing phase using clearly defined protocols for small scale aggregated trials linked to a PCM. The next section will outline the three stages of the model.

2.8.1 The planning stage

The first stage of the model, as shown in Figure 6 is the formulation of the research questions by a researcher and it is fundamentally important to involve practitioners during the process to ensure they understand why the research is needed and make a contribution to its development and feasibility from a practitioners perspective. The research design then informs two protocols for the pilot stage of the PCM. These are classified as the trial protocol and the implementation protocol. The purpose of a trial protocol is clear; to describe the rationale, methods, analysis plan and administrative details from the inception of the trial to the dissemination of the findings. The importance of clear and transparent protocols are fundamental requirements for all trials, as various stakeholders from researchers, ethics committees, funding bodies, journal editors and external reviewers rely on the validity of the information reported. The significance of this is supported by a growing evidence in clinical trial literature for the variation in the quality of trial protocols (Chan et al., 2004; Dwan et al., 2008; Pildal et al., 2005; Smyth et al., 2011; Tetzlaff et al., 2012).The consequences of poor or lack of

reporting can result in bias in the final trial results and unfortunately, educational RCTs are not required to produce protocols and publish these unless directed by funding providers such as the EEF. The first protocol involves the design of the aggregated trials, based on a checklist of best practice adapted from the SPIRIT 2013 guidelines in health. The second protocol is designed with the input of practitioners on how best to implement the intervention in the school. The involvement of practitioners will provide current insights into the logistical barriers which can impact on how an intervention is delivered in a school setting. Providing clear and concise easy to follow steps increases the likelihood that the intervention will be delivered as intended. If this is written in by a teacher, it should be easier for a teacher to follow. The advantage of developing protocols for a PCM is that the trial is repeated using the same criteria; such as eligibility, academic content, assessment materials, duration and time of delivery of the interventions, support available from evaluators, and so on. In traditional meta-analyses, these variables are often different and can be difficult to combine. The creation of a PCM should reduce the variation between trials when the results are combined.

The final section of the planning stage is the initial recruitment of schools for the pilot and the development of study instruments. Standardised assessments should be used as research demonstrates that researcher developed assessments often introduce bias and inflated effect sizes. However, if the trial is pragmatic driven with limited funds then unstandardised materials could be used and evaluated to determine their reliability and validity.

2.8.2 Implementation and testing

Each PCM will start with a pilot study to test the implementation protocol and instruments work as intended in the design of the study. The researcher will work with the practitioner to ensure that the study is implemented as intended and if small modifications are required, these are recorded in the protocol explaining why deviations have occurred. Randomisation should be conducted independently of schools by the lead researcher to reduce the risk of selection bias. Depending upon the design and type of intervention being evaluated, randomisation would be at cluster (school) or pupil (individual) level. For example, in the Every Child Counts intervention (Torgerson et al., 2011) the intervention consisted of daily 30 minute one-to-one session for target children over a period of 12 weeks in mathematics. Individual randomisation within each school with a waiting list design ensured all pupils received the intervention, ensuring all pupils requiring the intervention received this in the academic year. The individual randomisation of 12 children into three terms of delivery (autumn, spring or summer) allowed the effectiveness of the intervention to be assessed, as the post-test data from the children assigned to the spring and summer acted as controls to the children receiving the intervention in the autumn term.

However, some educational interventions cannot be delivered using individual randomisation, as these may involve whole school approaches. In these instances, schools recruited to the aggregate trial would then be randomised at cluster level to receive the intervention or act as a control. It is also possible to randomise within the school at teacher level if there is more than one class in the cohort, yet this could introduce threats to the internal validity of the study if the intervention through social interaction threats. Once randomised, the schools implement the intervention following the guidance set out in the protocols. As with previous aggregated trials in education (Gorard et al. 2016, Siddiqui et al. 2015) a light touch process evaluation will be managed by the researcher throughout the duration of the trial. As a protocol is used, a light touch process evaluation should be enough and by using less resources the cost associated with conducting the study should be smaller than traditional robust process evaluations. Finally, the analysis is completed as outlined by a specified analysis plan (SAP) and reports created for the schools participating. The data from cohort 1 is the first wave of schools in the PCM with an initial effect size reported.

2.8.3 PCM evidence base

Stage 3 is the replication of the aggregated trials for additional cohorts of schools, with the data accumulating in the prospective meta-analysis. An important question to ask is when does the data collection stop in a PCM? It could be argued that data collection should stop when sufficiency is reached, when effect sizes are consistently around a similar range. Consequently, once sufficiency is reached a new PCM is could be created to compare two effective interventions to determine what works best or the intervention can be scaled into a large RCT. Additionally, the PCM could extend into different sample populations within schools or focus on various academic subjects. For example, if the intervention involved one to one tuition in mathematics targeting students aged 11 – 12 years, an additional PCM could be designed for one to one tuition for student's aged 15-16 years. This would strengthen the validity that one to one tuition is equally effective for different age ranges if the effect sizes were consistent between the PCM's.

Finally, the data collected through the aggregated trials linked to a PCM cannot be used for sub group analyses not specified in the original protocols or SAP. Transparency is fundamentally important, therefore before collected data is released for further analysis a new protocol would be required and published to outline the research objectives in advance.

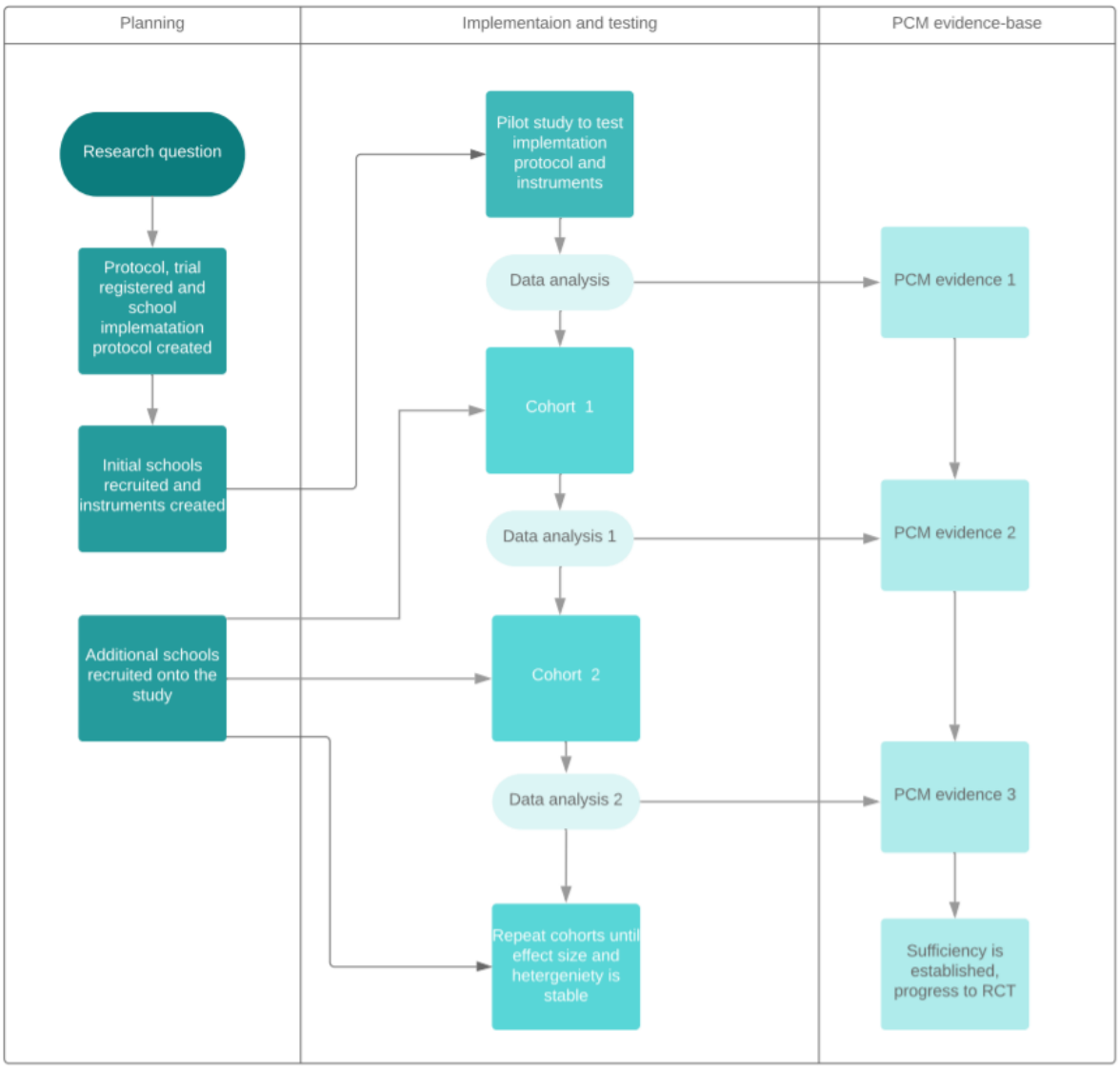


Figure 7 provides a model for the use of a standardised protocol for the aggregation small scale RCTs to create a PCM evidence base for a specific intervention in the testing phase of the research cycle.

The model can also be used to test the stability of an intervention after a large scale RCT is completed and the concept is explored further in the final chapter of this thesis.

2.9 Chapter conclusion

To summarise, evidence-based practice is the support and promotion of practices and policies that are based on good evidence about their effects (Coe et al., 2000). Through the implication that evidence-based education involves interventions, such as providing advice, implementing policies or promoting specific practices, imply that the purpose of this intervention is to bring about change.

Furthermore, when considering the effectiveness of an intervention, it is crucial to understand the counterfactual or the comparison being made. Higgins (2017) correctly suggests that the different kinds of comparison, through different research designs, provide evidence for stronger or weaker arguments about the robustness of any casual claims relating to the effectiveness of an educational intervention. For example, a randomised controlled trial has significantly less threats to its internal validity compared to a single group time series design. Chapter 3 methods section will outline the logic of appropriateness for the use of randomised controlled trials in the primary studies as part of the methods section of this thesis.

The increase in the use of randomised controlled trials, systematic reviews and meta-analyses has seen a global increase in the use of evidence to inform policy and practice. As outlined by figure 6, the full research cycle of social science research and development (Gorard, 2013) includes several distinct phases in the process, starting with an evidence synthesis and finishing with the dissemination, impact and monitoring phase of the cycle. Hargreaves (1996) suggested a gap existed between research and practitioners, between the rigorous testing and dissemination phase of the educational research cycle proposed by Gorard (2013). The empirical evidence clearly shows that the gap is still evident (Cousins & Walker, 2000; Lysenko et al., 2014; Nelson & O'Beime, 2014; Williams & Coles, 2007). This chapter has outlined the limited empirical evidence and potential barriers associated with the limited use of research by practitioners in schools.

The use of aggregated trials in education and the development of the prospective cumulative meta-analysis in healthcare provided the basis for the introductory model for evidence-based practice in schools using aggregated trials linked to a prospective meta-analysis. Through the use of pre-defined protocols, small scale aggregated trials conducted over a sequential time period allows the use of a prospective cumulative meta-analysis. The methodological advantages and limitations of a prospect cumulative meta-analysis will be discussed further in Chapter 9.

The next chapter of the thesis Chapter 3, introducing the methodology for the two primary research studies involved in the development of a PCM.

Chapter 3: Methodology

This chapter outlines the research method for the two primary studies involved in the development of small-scale aggregated trials linked to a prospective cumulative meta-analysis. The first section outlines the logic of appropriateness for the use of randomised controlled trials as the preferred methodology, specifically linking to internal and external validity. The issue of validity applies to both the studies for online peer tuition and online small group teaching involved in this thesis, therefore it is important to understand the logic of appropriateness by discussing these concepts before outlining the methodology. The second section outlines the methodology for the first study for the pilot efficacy trial for online cross-age peer tuition between primary and secondary schools, exploring the effectiveness of online group sizes 1:1, 1:2 and 1:4. The third section outlines the methodology for the second study investigating the effectiveness of online small group teaching of year 7 pupils (aged 11 – 12 years) for the mathematics topic of fractions.

In advance of the chapter commencing, it is useful to outline the sections to demonstrate the logic for the structure. Table 2 provides and overview of the key sections and how these link to the development of small scale randomised controlled trials linked to a prospective cumulative meta-analysis.

Table 2: Overview of studies

Section	Reasoning
Internal and external validity	In order to understand the logic of appropriateness for using RCT methodology for the aggregated trials, internal and external validity will be discussed.
Logic of appropriateness	The reasoning for selecting RCT methodology for the design of the two research studies will be outlined.
Chapter 4	Pilot efficacy study for the small-scale aggregated trials investigating the effectiveness of online peer tuition group sizes.
Chapter 6	A feasibility study to develop a prospective cumulative meta-analysis based on aggregated small scale RCT to evaluate the effectiveness of a TUTE mathematics intervention.

3.1 Internal and external validity

Validity is the term used to refer to the approximate truth of an inference (Shadish, Cook & Campbell, 2002). If we infer something is valid, a judgement is made on the extent of which the evidence supports the inference as being true. It is important to note that validity is a property of inferences and not a property of designs or research methods. It is not correct to assume that by selecting an RCT you are guaranteed to be able to generate valid inferences, as many threats can occur through poor implementation and planning. Ginsburg & Smith, (2016) contend that poor implementation fidelity, non-independent evaluations and assessments designed by the developers of the intervention can provide threats to validity, resulting in inferences which could be partly or completely incorrect. Before discussing why an experimental design was considered the most appropriate research design for the primary studies, it is important to consider the key terms internal and external validity.

3.1.1 Internal validity

Internal validity is defined by Shadish et al. (2002) as "the validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured". As internal validity involves inferences about causal relationships, Trochim (2006) states that this is only relevant to studies that try to establish this causal mechanism. If the research questions require observational studies or are descriptive, then internal validity is not a primary concern. In order to determine if an intervention is successful a change must be made, rather than simply observing or describing existing situations. Therefore internal invalidity is a key consideration for most educational research studies which seek to address efficacy or effectiveness questions (most education RQs are not), as the fundamental question is whether the observed changes (such as test scores, increased pupil attendance, increased participation in STEM subjects) can be attributed to the intervention and not to other plausible explanations.

3.1.2 External validity

External validity is the degree to which the conclusions drawn from the study would hold for other persons in other places and at other times (Trochim, 2006). Therefore the criticism that once an RCT is completed, the conclusions drawn from the study can be inferred onto the whole population is an area of contention in educational research. Deaton & Cartwright (2016) suggest that external validity may just refer to the transportability of the causal connection, or it may go further and require the replication of the average treatment effect. The authors argue that this is unhelpful, as it overstates

and understates the value of RCTS. For instance, this view directs us towards a simple extrapolation of the results - will the results hold elsewhere or a simple generalisation as to whether it holds universally. Deaton & Cartwright explain further that establishing causality through the use of RCTs does nothing in and to itself to guarantee generalisability.

Referring back to the initial criticism outlined by Hargreaves (1996) that educational research is non-cumulative in terms of a lack of replication of studies, this links into the argument on the external validity of RCTs. In order to strengthen the validity of trial results, replication of RCTs should be conducted and the results synthesised as part of a meta-analysis. It could be argued that if the results of a well conducted RCT, with a sufficient power analysis were to be replicated, this would be a waste of resources in terms of personal and cost. Deaton & Cartwright (2016) still argue that even with multiple replications we cannot provide much support or guarantee the conclusion that the next will work in the same way. However, if failures to replicate occur, these should not be classed failures as it allows researchers the opportunity to understand potential factors as to why this occurred. This is an important point, as education is a complex environment in which to run rigorous research studies at scale. Therefore, before we commit to scaling feasibility or efficacy studies into large scale RCTs, we need to understand through the use of protocols and aggregated trials linked to a PCM if the study has a stable effect size during the replication process.

Higgins (2017) highlights the importance of tracking fidelity and how faithfully those involved in the study adopted the changes to practice. In practice, it may not be possible to expect educational interventions to be implemented identically due to the social nature of education. Yet, as Cartwright (2013) explains that the appeal for fidelity is not to do the same thing in exactly the same, but rather to do something faithful to the higher level principle.

In terms of external validity of the effect sizes of RCTs in education, it may be useful to think of the average treatment effects in terms of distributions as well as averages. Deaton & Cartwright (2016) explain that the inferences taken from the means can be hazardous as in the presence of outliers, means themselves do not provide the basis for reliable inference. Through looking at the distributions around the effect size, rather than the average treatment effect allows practitioners a best bet approach. This just means that if they implement an intervention based on evidence from RCTs, that they are likely to see an improvement between the range of the distributions. Through the use of small scale aggregated RCTs linked to a prospective cumulative meta-analysis, it will be argued that this allows a viable mechanism for replication. In addition to increasing replication, the nature of the cumulative meta-analysis allows for the distributions to be analysed separately and cumulatively,

enabling researchers the opportunity to explore the fidelity of trials if the effects varied significantly from the initial large scale RCT.

The importance of considering the distributions as well as the average treatment effects will be discussed in further details in the discussion of this thesis. The next section of this chapter explores the appropriateness for using randomised controlled trials for the aggregated studies in the primary research for this thesis.

3.1.3 Logic of appropriateness for the trial designs

The evidence base of the effectiveness of online learning interventions often uses case study designs with small sample sizes, with Means et al., (2009) reporting very few experimental or quasi-experimental designs used for studies involving school aged children. Recent research from the EFF investigated the effectiveness of a one to one online tutoring programme where primary pupils received tuition from trained graduates in India and Sri Lanka (Torgerson et al., 2016). In order to demonstrate the logic of appropriateness for the two primary studies using online learning in this thesis, the Affordable Maths Tuition evaluated by the Torgerson et al., (2016) will be used as a model to explain the strengths and limitations of randomised controlled trials.

The study investigated the effectiveness of an affordable online maths tuition programme on the maths skills of participating children. A case study is not a suitable design as the research question is an effectiveness question with a causal inference, in that the online maths tuition has an impact on the participating children. Case studies have multiple threats to internal validity, as the lack of a comparator group results in many alterative plausible explanations for any observed improvements. These threats include maturation threats whereby the improvement in test scores might happen naturally even if the tuition did not occur (Trochim, 2006). For example, if pre and post testing are involved in assessing academic performance, testing threats whereby the pre-test can affect the post-test through familiarity of the questions could occur. In addition, regression to mean whereby extreme scores are likely to change in the post-test regardless of the intervention could also provide a plausible explanation. Finally, as case studies are single group designs, the pupils could also receive alternative interventions inside or outside of the school environment. Therefore history threats present a valid threat to the internal validity of this study design. Trochim (2006) states that RCTs have the ability to overcome all threats to a single-group study design except for social interaction threats and attrition.

The evaluators of the online maths intervention could use an ex facto or quasi-experimental design, as these use comparator groups and minimise the threats to the validity described in single group designs. However, this design is still susceptible to multiple-group threats. These include:

selection maturation threats, selection history threats, selection instrumentation threats, selection regression threats and selection testing threats. Bias can occur in the selection of the groups, as it is not possible to match groups for unseen hidden variables. RCTs overcome all multiple-group threats with the exception of attrition and social interaction threats.

The design used by Torgerson et al. (2016) to evaluate the effectiveness of the online intervention used a pragmatic cluster randomised controlled trial involving 64 schools and 600 pupils. The RCT design is a design which can be used to establish whether or not an intervention is effective (Shadish, Cook & Campbell, 2002; Torgerson & Torgerson, 2001). The design creates two or more groups by random allocation of individuals or clusters, ensuring that all unknown and known variables are equivalent at baseline with the exception of chance. Therefore the threats to the internal validity of single-group designs and multiple-group designs are reduced, with the exception of attrition and social interaction threats. Randomisation primarily deals with the allocation bias, yet other forms of bias can impact the validity of studies. If we look systematically, these can impact in the design, process and analysis of a study. Firstly, in the design, bias can be introduced through a lack of randomisation or outcome measurement. Secondly, in the process a lack of blinding can lead to ascertainment bias whereby people know which group they are in or high levels of attrition may create imbalances in the groups. For example, if differential attrition occurs this may vitiate the process of randomisation or the power may simply be too low. Finally, in the analysis selective reporting can also threaten he validity of the findings of the study.

In addition to these threats, poorly planned and executed RCTs can also present threats to the validity. As Torgerson (2009) succinctly explains, poorly designed and reported trials are not helpful to anyone and can easily mislead. The use of a trial protocol and statistical analysis plan is designed to address most of these concerns.

A common argument in the literature against the use of RCTs in education is that it is unethical to conduct experiments on children, as the use of a control group prevents all pupils from receiving the intervention (Morrison, 2001). If we know as researchers that a particular intervention has robust evidence of a positive impact, then the argument against the use of an RCT would hold. Yet, the reason RCTs are used is to increase our confidence in an intervention working, so it should be argued that it is unethical to not use an RCT if the evidence is not clear. This is a circular argument as we can only know something works through the use of an RCT.

Fitz-Gibbon (2004) explains two examples where good intentions actually harmed children. Scared Straight was a programme based on the theory that if children showing signs of delinquency were taken to prisons to be told by inmates how horrible prison was, they would be deterred from a

life of crime. Many subjective responses such as “it kept me away from crime” were given regarding the effectiveness of the programme, but through the use of a control group left untreated (randomised) it was found that the programme led to more serious crimes and increased recidivism in the treatment group compared to the control (McCord, 1978, 2001). A second example explains how the good intentions of providing counsellors to provide psychological debriefing immediately after a trauma and accidents. This seems a logical process to support patients’ wellbeing. However, Fitz-Gibbon (2004) highlights two randomised trials which concluded that this will do more harm (Mayou, 2000; Wessely, S., Rose, S., Bisson, 1998).

Furthermore, as Davies (1999) points out that the ethical questions as to whether or not it is right or warrantable to undertake the intervention must be considered. Any type of research, regardless of the methodological approach used will require selective action, use informed choices, require resource and time allocation. In educational research, it is important to establish the need for the research and then use the most appropriate methodology to answer the research questions.

In Study One for the online cross age peer tuition trial, the use of an active control through the 1:1 group size ensured all pupils received the intervention. All pupils receive the same intervention, only the group size changes from 1:1, 1:2 and 1:4. The ethical issues raised by (Morrison, 2001) are removed, as all children receive the intervention at the same time. In addition, the use of an active control strengthens the design by minimising the threat of social interactions as all pupils are using the technology to receive the intervention. Therefore threats to validity for compensatory equalisation, selection mortality, demoralisation and rivalry are reduced.

Study Two investigated the effectiveness of online group teaching, the trial uses a standard control group whereby the intervention is withheld. In order to counter the ethical concerns raised by Morrison (2001), the addition of a second stage with the wait list was used and reduces the impact of social interaction threats which can occur in RCTs (Slavin, 2004; Thurston & Topping, 2008). At the start of the trial, all schools, parents and pupils were informed that if they were not selected for the intervention in the first instance, it would be offered to them if the intervention was effective after the study had been completed.

In sum, as the research questions for the two primary research studies in this thesis are investigating the effectiveness of online learning, through online peer tutoring and small group teaching, the most appropriate research design is a randomised controlled trial in order to make causal claims. The following methodology outlines how the trials are designed to minimise threats to the internal validity.

3.1.4 The development of trial protocols

The importance of trial protocols was not fully understood until the first cohort for the primary Study One had begun. Appendix 1 includes the limited ethics proposal and trial diagram submitted to Durham University Ethics Committee prior to the start of the research. As the importance of structured protocols and statistical analysis plans became apparent, these were written prior to the second cohort of schools completing the research. These are included in Appendix 2 and 3.

Chapter 8 of this thesis explores the importance of protocols, cumulating in a proposed protocol checklist for conducting research in the social sciences.

Chapter 4: Pilot efficacy study for small scale aggregated trials investigating the effectiveness of online cross-age peer tuition group sizes across the transition boundary between primary and secondary schools (Study 1).

The first study builds on the previous MA dissertation titled "A Feasibility Study: To determine the feasibility of implementing a mathematics online cross-age peer tuition intervention across the transition boundary between primary and secondary schools" (Harrison, 2015). The study successfully implemented a small scale randomised controlled trial using secondary school year 7 pupils to peer teach mathematics online to primary school year 5 pupils. However, the limited sample included only 4 schools and 38 pupils. Furthermore, as the intervention required direct monitoring and support for each online peer tuition session, the feasibility of conducting one larger randomised controlled trial for the PhD study was not possible due to a number of restrictions. Firstly, each individual partnership between a primary and secondary school required planning involving the co-ordination of the delivery, assessment and IT testing with two IT technicians. Secondly, school firewall systems and computers often required troubleshooting before some peer tuition sessions were previously delivered (Harrison, 2015). This required the researcher to be present to observe the implementation and delivery of the online intervention. Thirdly, due to the limited resources, such as financial (travelling expenses) and only one researcher managing all aspects of the study, it would not be feasible to oversee such an intensive programme if delivered at scale.

Consequently, through the use of planned small-scale aggregated trials across a sequential time period using a specific protocol to enable the replication of the individual studies, it should be possible to reduce these logistical barriers. The lessons learned from conducting the small-scale studies provide valuable insights in informing future large-scale studies, potentially saving time and resources. However, as discussed in the limitations in chapter 10, due to the intensive nature of the planned intervention the sample size will still be under-powered. Therefore the primary study has been defined as a pilot efficacy study using small scale aggregated trials linked to the development of a cumulative meta-analysis.

4.1 The intervention and rationale for conducting the research

This section will start with outlining the rationale and intervention for completing the study, then moving onto the traditional methods reporting for an RCT.

4.1.1 Rationale

The transition from primary to secondary school in the UK has been highlighted in a recent Ofsted report entitled 'KS3: The wasted years' as a key area for improvement for schools. A weakness identified in numerous schools is the lack of an academic focus for more able and talented pupils in UK schools during the transition from primary to secondary schools (Ofsted, 2015). Transition between primary and secondary schools has been depicted as 'one of the most difficult phases in pupils educational careers' (West et al., 2010). Universal agreement exists in the evidence that the majority of pupils express some concerns and anxieties prior to the transition from a primary to secondary school. Issues identified in the research range from the size of the school to peer relations (Chedzoy & Burden, 2005; Graham & Hill, 2003; Shepherd & Roker, 2005). A long term longitudinal study investigating pupils experiences of the primary-secondary school transition in the West of Scotland demonstrated the importance of a successful transition on later well-being and attainment (West et al., 2010).

The Education Endowment Foundation teaching and learning toolkit (Higgins et al., 2015) highlights peer tuition as a cost effective intervention with suggested gains of +5 months progress. The benefits are apparent for both the tutor and tutee, particularly in cross-age peer tuition. Despite the robust evidence supporting the academic impact of peer tuition, recent surveys by the National Audit Office (2015) and Sutton Trust (2017) find few schools who are using this intervention, particularly as a transition intervention between primary and secondary schools. Multiple reasons exist in schools for not using this effective strategy, these include logistical barriers such as the cost of transport, health and safety administration, organising rooms (rescheduling other classes), time taken to travel between schools and the pressures of repeating the process over a sustained period of time (Harrison, 2015). In addition to the logistical barriers, secondary schools will often work closely with over 5 feeder primary schools, requiring any intervention to be potentially offered to each school. Therefore traditional peer tuition as a transition intervention is often not feasible.

Harrison (2015) successfully implemented a feasibility study for online cross-age peer tuition across the transition boundary between primary and secondary schools. Therefore, the current study follows on from the initial research to implement the intervention on a larger scale through the development of small scale RCTs linked to a prospective cumulative meta-analysis. Furthermore, if one-to-two and one-to-many (1:4) can be as effective as one-to-one online peer tuition, secondary schools can use small numbers of peer tutors to deliver a flexible and cost-effective transition intervention to feeder primary schools.

4.1.1.1 What is the existing evidence on using peer tuition to raise academic attainment?

The evidence on the academic impact of peer tuition is fairly robust, with a number of meta-analyses providing evidence that peer tutoring positively impacts tutee achievement (Cohen, Kulik, & Kulik, 1982; Cook et al., 1985; Ginsburg-Block, Rohrbeck, & Fantuzzo, 2006; Leung, 2015). Table 3 indicates that the median indicative effect size for peer tutoring is +0.43, with the number of studies within each meta-analysis ranging from 11 to 72.

Table 3: Summary of effects of peer tuition from meta-analyses (adapted from EEF, 2015)

Study	Effect Size
Cohen, Kulik and Kulik, 1982 (tutees)	0.40
Cohen, Kulik and Kulik, 1982 (peer tutors)	0.33
Cook et al. 1985 (tutees)	0.59
Cook et al. 1985 (peer tutors)	0.65
Rohrbeck et al., 2003	0.59
Ginsburg-Block et al., 2006	0.35
Jun, Ramiz & Cumming, 2010	1.05
Bowman-Perrot et al., 2013	0.75[3]
Washington State Institute, 2014	0.43
Leung, 2015	0.26
Zeneli et al. 2016	0.21
Indicative effect size (median)	**0.43**

Cohen et al. (1982) identified subject areas, intervention features (grade and ability), duration and student characteristics as having a moderating effect on peer tuition. It is important to note that this meta-analysis preceded the development of many recent methodological advances currently used in these studies today. The study does not include homogeneity analyses, fails to correct for small sample sizes, weight effect sizes by sample sizes or consider sample size outliers. In addition, the study combines both adult and teacher led tuition, which has the potential to confound the results.

The methodological limitations are not restricted to this one study, with a number of recent meta-analyses which indicate a positive effect for peer tutoring (Cook et al., 1985; Mathes & Fuchs,

[3] 26 single case research studies using TauU (CI 0.71 – 0.78), a non-parametric effect size indicating moderate to large effects.

1994; Rohrbeck et al., 2003) with methodological design limitations. For example, the early meta-analysis of Cook et al. (1985) has similar methodological issues to the Cohen et al. (1982) meta-analysis, as they fail to correct for small sample sizes, fail to conduct statistical heterogeneity analyses for random or fixed effects and do not use weighting procedures for effect sizes. Rohrbeck et al. (2003) combines peer tutoring with other forms of learning, whilst failing to use any procedures to address publication bias or reporting estimates for both fixed and mixed effect models for the effect size.

However, two recent meta-analyses (Leung, 2015; Zeneli et al., 2016) investigating the moderators and crucial determinants of the effectiveness of peer tuition and the influence of experimental design on the magnitude of effect size have demonstrated that peer tutoring has a moderate positive impact on academic achievement. These studies address the methodological limitations of earlier studies, with effect sizes of 0.26 (Leung, 2015) and 0.21 (Zeneli et al. 2016).

Evidence from recent RCT studies on peer tutoring have reported lower effect sizes than the median indicative effect size of +0.43 in table 2.7. A large scale RCT was conducted for the Fife Peer Learning project involving peer tutoring for reading and mathematics, reported a mean effect size of + 0.20 – 0.35 (Thurston & Topping, 2008; Topping, Miller, Murray, & Conlin, 2011; Tymms et al., 2011). In addition to the Fife studies, the Education Endowment Foundation recently evaluated two large scale RCT studies investigating the impact of peer tuition, Durham Shared Maths Project (Lloyd et al. 2015a) and Paired Reading (Lloyd et al. 2015b). Both these studies found minimal effect sizes for the impact of peer tuition.

The moderator effect of the research design is highlighted by Leung (2016) and Zeneli (2016), whereby peer tuition studies implementing RCTs have a smaller effect size than studies implementing matched quasi-experimental designs. The potential issue of the impact of research designs on effect sizes is discussed in detail in chapter 5, however it is important to recognise this point when considering the current evidence.

4.1.1.2 What is the existing evidence on the impact of digital learning on academic attainment?

The evidence base for the impact of digital learning on academic attainment is currently weak, with the majority of online learning studies plagued with theoretical and methodological flaws. Means et al., (2009) conducted a meta-analysis to evaluate evidence-based practices in online learning and found no experimental or quasi-experimental designs comparing face to face instruction and online learning with school aged pupils. A systematic search of the literature between 1994 to 2006 found no studies with experimental designs. The lack of robust studies in the final meta-analysis resulted in the authors combining different types of learners: school children, college level, university level and professional development for a sample of 50 studies, with only 7 involving children under the age of

13. In addition to the limited number of studies with experimental designs, only 5 studies from the sample included more than 400 learners. Taking the limitations into account, the study reported that students in online conditions performed better on average than those learning the same material through traditional face-to-face instruction with an average effect size of +0.20.

A synthesis of 45 meta-analyses on the impact of digital technology on academic attainment (Higgins, Xiao, & Katsipataki, 2012) from 1990 to 2012 found the average effect size to be between 0.3 and 0.4. However, the authors acknowledge a wide range of effect sizes, from -0.03 to +1.05 and suggest that it is not whether or not technology is used to make a difference, but how well the technology is used to support teaching and learning.

Finally, it is important to recognise that the constant and rapid advances in technology make it difficult to create a robust evidence base (Falloon, 2014; Higgins et al., 2012). However, it is crucial that researchers use the most appropriate research designs for evaluating the effectiveness of interventions. Through the use of small scale RCTs linked to a prospective cumulative meta-analysis, robust evaluations on the impact are potentially possible.

4.1.1.3 What is the existing evidence on the impact of online peer tuition on academic attainment?

The evidence base for the impact of online peer tuition on academic attainment is similar in quality to the evidence for the impact of digital technology, limited with methodological and theoretical flaws. Online peer tuition can be categorised as either asynchronous or synchronous communication. The term asynchronous refers to computer-mediated communication (CMC) which is time delayed, therefore not relying on simultaneous access for educational outcomes (Oztok et al., 2013). Synchronous refers to computer-mediated communication in real time between learners (Tsuei, 2012).

The evidence on the impact of asynchronous peer tuition is mixed on whether this type of learning has a positive impact on students. Rourke & Kanuka (2009) suggest that the vast majority of posts written by students in asynchronous platforms fail to provide a cognitive challenge. Tu & Corry (2003) argue that this form of communication can lead to in-depth discussions, with this view supported by Branon & Essex (2001).

Topping et al. (2013) investigated the paradoxical effects of feedback in international online reciprocal peer tuition using asynchronous feedback. The study found that the reading comprehension and writing tasks indicated a significant improvement in the students who were aged 9 – 12 years

from Scotland and Catalonia. This supported the previous research by (Dekhinet et al. 2008; Thurston et al. 2009) in the use of online reciprocal peer tuition for international students.

The evidence on the quality of synchronous versus asynchronous delivery on learning is questionable (Hrastinski, 2006; Rockinson-Szapkiw, 2009; Schwier & Balbar, 2002) due to the research designs used in the studies. Case studies appear to be the most common research design used in the literature, with studies often using convenience sampling (Rockinson-Szapkiw, 2009). In addition to the inappropriate research designs and the lack of comparator groups, studies often use unstandardised and researcher made assessments. The use of unstandardised assessments is linked to inflated effect sizes and can introduce bias into the study. For example, a study by Kuyath (2008) used grades on an assignment as the outcome measure and suggested this as evidence that synchronous chat can improve academic performance.

Tsuei (2012) used a synchronous peer tutoring platform to enable pupils aged 10-11 year to peer tutor in mathematics, with the experimental group having significant learning gains compared to the control. The study included 88 children and used a quasi-experimental design, with two classes assigned to the experimental group and one class assigned as the control. The study demonstrated positive effects on the students self-concept and attitude towards mathematics, through the use of peer support on the online synchronous learning platform (Tsuei, 2011).

Ruane (2012) explored the student interaction in an online learning environment specially crafted for cross-level peer mentoring. The research focused on social interactions of users over a 10 week online module, rather than the impact on academic attainment of the students.

Lin & Yang (2013) explored college students' experiences and perceptions of using google documents and online English 4th year majors for the cross-age peer tutoring of forty-four 1st year non-English majors from a college in Taiwan. However, the research focused on attitudinal survey responses, yet the authors stated that online tutoring enhanced the students English writing skills.

Kivunja (2015) studied the attitudes, characteristics and knowledge of learners linked to peer mentoring. The sample included 145 2nd year university education students using a case study design. As the author states "convenience sampling procedure" was used in selecting the participants because of the ease of the researcher's access to these cohorts of students and their willingness to participate in the study". Design limitations in the research are a recurring theme in the evidence base for online learning, with a number of recent studies research designs focusing on the attitudinal or user journey rather than effectiveness of the intervention.

In sum, the research evidence for traditional peer tuition is based on a fairly robust evidence base with an indicative effect size of +0.43. Evidence from recent meta-analyses (Kim Chau Leung, 2015; Zeneli et al., 2016) suggests that the type of research design has a moderating effect on the effect size of peer tutoring interventions. The evidence base of the impact of digital technology and online peer tuition is weak, with very few studies able to meet the requirements for a meta-analysis. It is not possible due to the limitations of this research to conduct a systematic review of the literature to assess the quality of evidence for online learning, however this will be discussed in greater depth later in this thesis.

4.1.2 Intervention

The intervention was designed to allow Year 8 secondary pupils (12-13 years) to peer tutor Year 6 pupils (10 – 11 years) on the topic of data handling in mathematics. The topic was chosen through discussions with the initial feeder primary schools identified for the pilot study, as this is often an area of weakness in mathematics. The lessons are delivered online by the Year 8 peer tutors for 30 minutes for 5 weeks, using the TUTE online learning platform.

Prior to the start of the intervention, the dates and times were co-ordinated and agreed between the participating primary and secondary schools. The aggregated trial includes three cohorts; Cohort 1 delivered the tuition between November – December 2015, Cohort 2 delivered the tuition between February – March 2016 and Cohort 3 delivered the tuition between June – July 2016. Table 4 provides an overview of the schools and the delivery times for the online peer tuition.

Table 4: School delivery of the online peer tuition for the aggregated trial

School (Secondary)	School (Primary)	Cohort	Delivery time
School A	School B	1	Tuesday 9-9.30am
School A	School C	1	Wednesday 9-9.30am Thursday 9 – 9.30am
School D	School E	1	IT issue so unable to deliver (retrospective control)**
School F	School G	2	Tuesday 3.30 – 4pm Wednesday 3.30-4pm
School A	School H	3	Tuesday 9-9.30am
School A	School I	3	Wednesday 9-9.30am
School J	School K	3	Thursday 10.30– 11am Friday 11.30 – 12pm*

** Retrospective control is a control group which was not originally planned in the study or SAP. Results of the analysis are only included in the appendix as this is outside the scope of the original research.

*The school requested an alternating timetable so that the pupils did not miss the same lesson each week.

However, before the intervention could be delivered in schools a schedule was created to strategically plan the trials for each cohort. For example, the project plan includes the initial meeting between the partner schools, ethics approval, IT testing with each schools nominated technician to open the relevant firewalls, pre-assessment for tutors and tutees, tutor training, 5 x lesson observations, post-assessment for tutor and tutees, focus groups and interviews with key personnel in schools.

Appendix 2 provides a detailed protocol for the three cohorts involved in the aggregated trial.

4.1.2.1 Online learning platform

The online learning platform used by TUTE incorporates the BigBlueButton online software as the collaborative learning environment. The BigBlueButton uses a flash-based software, restricting the delivery to the use of school-based computers, as tablets were not compatible with this technology. This issue will be further discussed in the process evaluation as it created a logistical issue with timetabling computer rooms in some schools.

The platform incorporates video link for the teacher (this was disabled for peer tutors), a chat room for multiple text from students and teachers, active users in the virtual classroom with the ability of teachers to mute students and a collaborative whiteboard area. The peer tutor uploaded a PDF of the lesson prior to the start of the lesson, with the ability to annotate the lesson resources and switch this function to individual students to annotate questions.

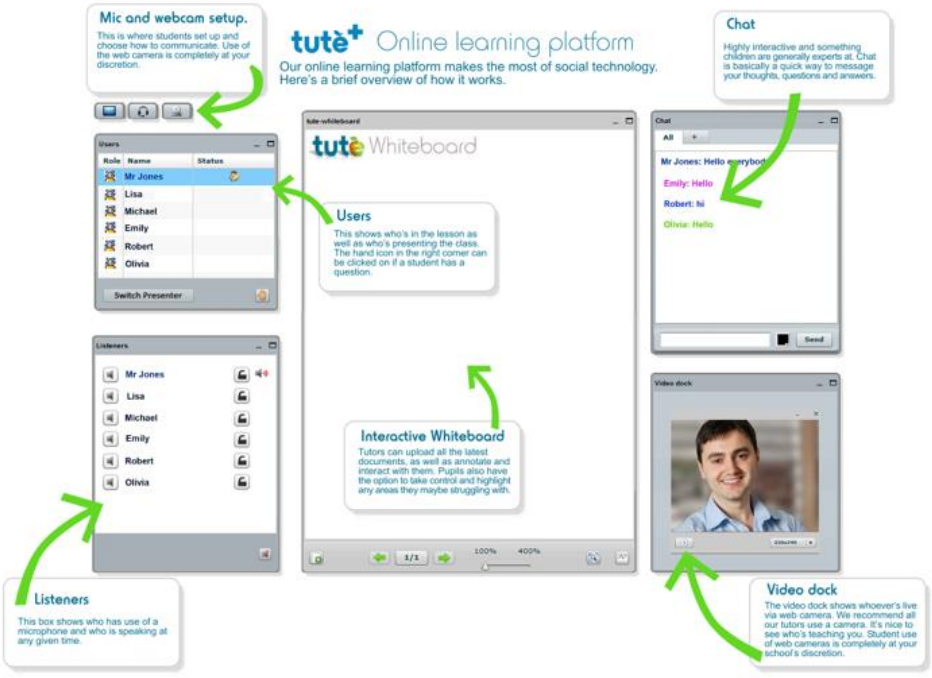


Figure 8: Visual representation of the online learning platform features.

The choice of the online learning platform involved three potential platforms; Microsoft 365 Lync, Google Hangouts and the TUTE online Learning platform. However, after testing the three platforms it was only possible to restrict user access through the TUTE platform, restricting communication to

the booked lesson time with all lessons recorded as a child protection feature. It was not possible to restrict online communication between the tutor and tutee on Microsoft 365 or Google Hangout, presenting a safeguarding issue with schools. Consequently, the TUTE online platform was selected as the learning platform for the delivery of the online peer tuition.

4.1.2.2 Lesson resources

The online peer tuition lessons were designed by the researcher in partnership with a year 6 teacher from school B and checked by an independent maths teacher at Tute. The five online lessons focused on data handling, including an introduction lesson, interpreting bar graphs, pie charts, scatter graphs and a revision lesson. The purpose of the introduction lesson was to allow the peer tutor and tutee to familiarise themselves with the learning platform, set ground rules for positive behaviour and discuss transition experiences.

The lessons were printed into handouts and provided to all the peer tutors one week in advance of the peer tuition, allowing the tutors an opportunity to become familiar with the lesson content prior to delivery.

Each lesson included specific learning outcomes linked to tasks, with the tutees able to rate their progress on a scale of 1 – 5 in the chat box. Tutors were able to revisit a learning outcome if their tutee struggled to grasp the concept, however this created issues if the online group sizes were 1:4. Figure 9 demonstrates an example of a lesson resource from the intervention.

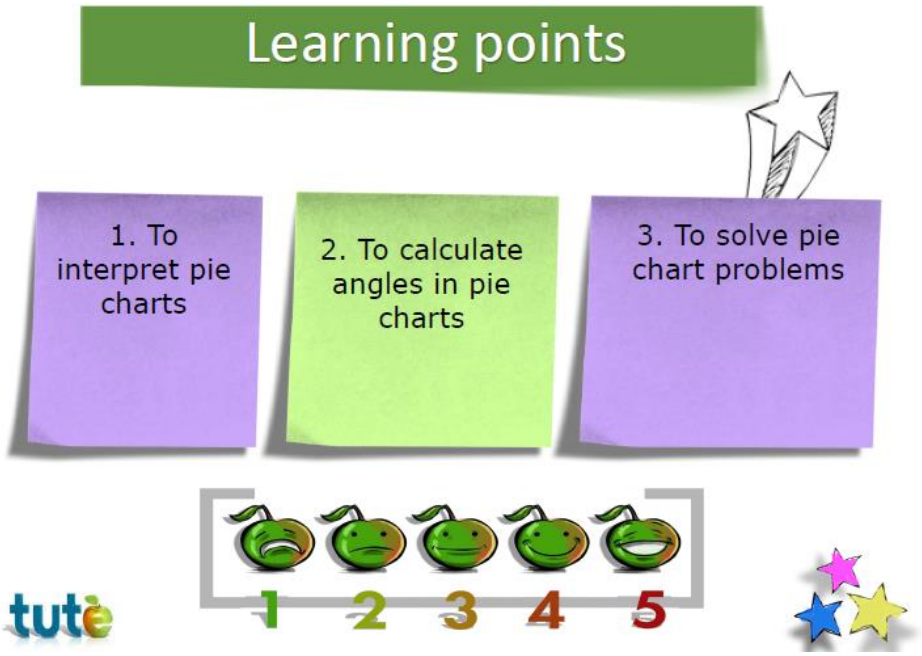


Figure 9: Example lesson resource

The structure of the lessons involved peer tutors modelling mathematical questions and providing support in answering the questions if this was required. Figure 10 shows how the mathematical problems were presented to the tutees, with Figures 11 and 12 demonstrating how the tuition was scaffolded for the peer tutor to provide answers.

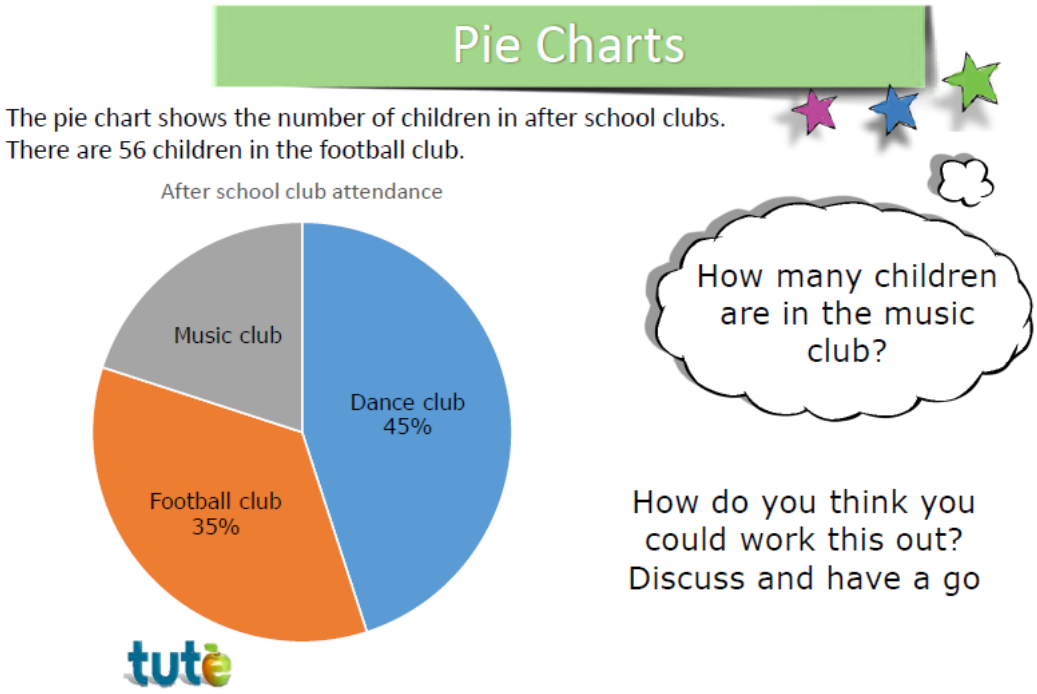


Figure 10: Pie chart example question

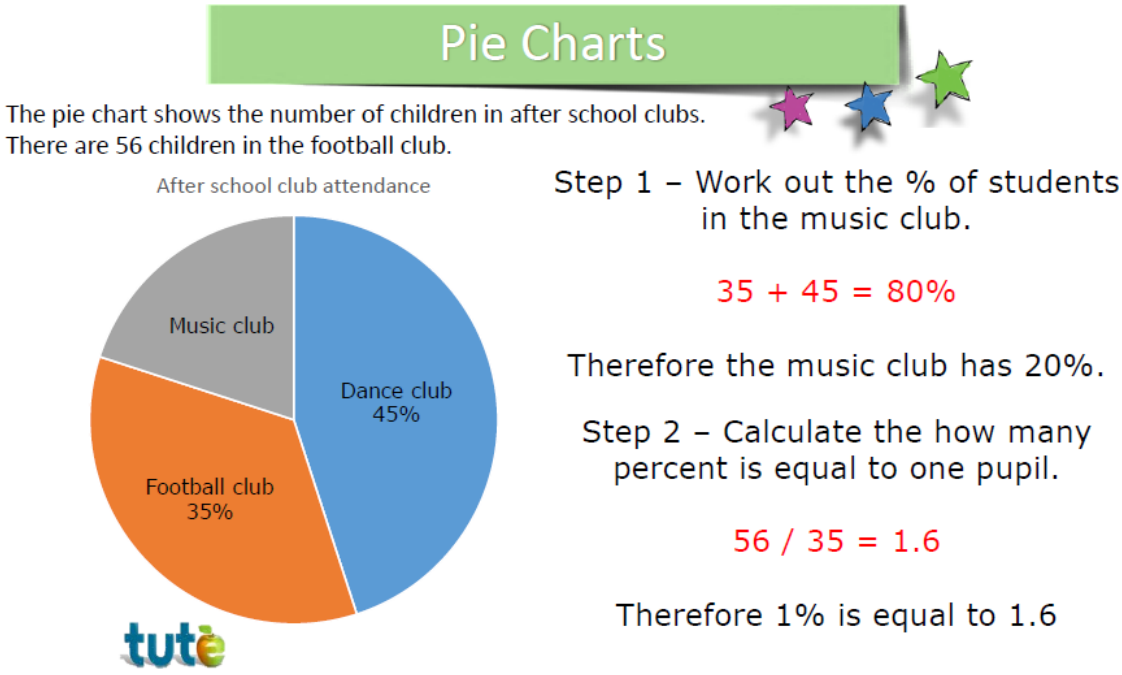


Figure 11: Pie chart example of scaffolded support on how to answer the question

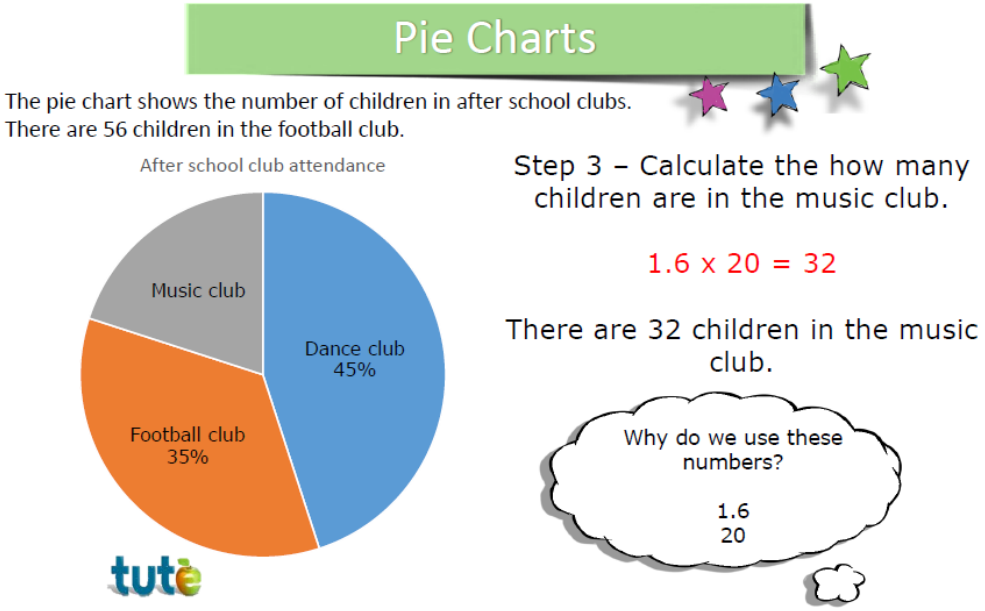


Figure 12: Pie chart example of scaffolded support on how to answer the question

The students revisited the intended learning outcomes at the end of the lesson and extension materials were provided if these were required by the peer tutor. The teacher involved in the content creation advised the extension material to be based on applying and mathematical reasoning.

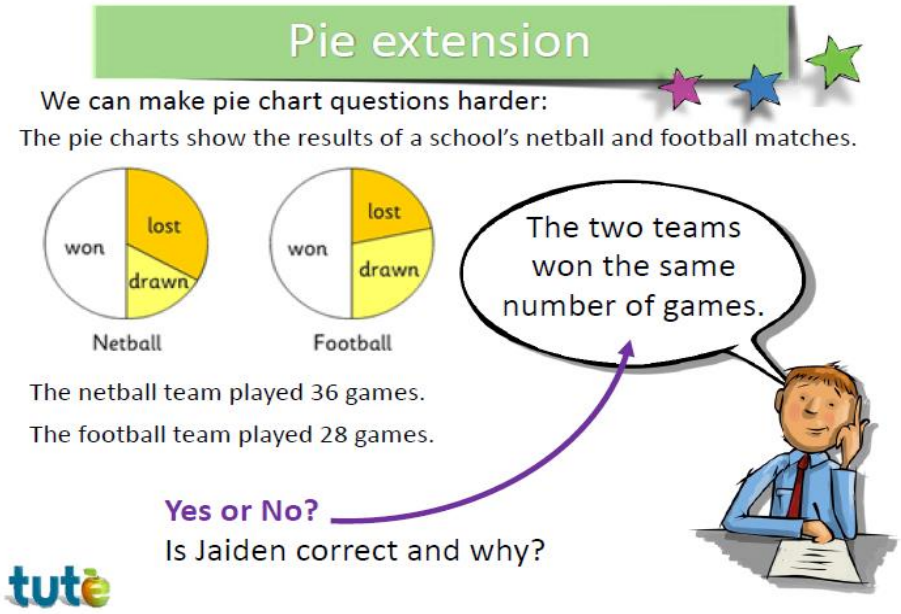


Figure 13: Extension material for the peer tuition lessons

4.3 Methodology

4.3.1 Research questions

The objective of the study was to determine the feasibility of implementing a more able and talented cross-age peer tuition intervention across the transition boundary between primary and secondary schools. The principal research questions are:

1) Is the implementation of an online cross-age peer tuition intervention feasible within primary and secondary schools as an aggregated trial?

    a) Is online cross-age peer tuition a feasible intervention strategy within schools as part of an aggregated trial?
    b) Are the instruments used appropriate for the ability of students involved in the study?
    c) Can the intervention maintain fidelity if delivered to schools in other locations in the UK?

2) Does one-to-two online cross-age peer-tutoring produce equivalent effects in mathematics attainment[4] compared with online one-to-one tuition?
3) Does one-to-many online cross-age peer-tutoring produce equivalent effects in mathematics attainment[2] as compared with one-to-one tuition?

4.3.2 Trial Design

The design for the study is a pragmatic randomised controlled trial using an aggregation of three smaller trials over a sequential timescale of 9 months. The individual trials were conducted as an efficacy study as the intervention was required to be conducted under highly controlled and optimal conditions in order to maximise outcomes (Wiglesworth et al., 2016), and ensure the technology worked in schools. Recruitment targeted secondary schools with feeder primary schools in the North East of England, recruiting 7 Year 8 peer tutors (12 – 13 years) to tutor 12 Year 6 pupils (10-11 year) per school partnership. The Year 8 pupils were randomised using unequal randomisation, controlling for any unknown hidden biases when conducting the final analyses. This is a limitation in the research as sequence generation would reduce the risk of bias during this process. The Year 6 pupils were block randomised using pre-test data ensuring the active control and two comparator groups are balanced for the mathematics ability as defined by the data handling assessment. The design is ethical as all pupils are participating in the intervention, with only the group sizes changing for the delivery of the

[4] Mathematics attainment is defined as mastery level concepts in KS2 curriculum for Algebra.

peer tuition in group sizes 1:1, 1:2 and 1:4. The use of an active control ensured that all pupils involved in the intervention used technology for the delivery of the intervention, reducing the potential impact of social interaction threats to internal validity and reducing the impact of the Hawthorne effect.

The research design is illustrated using the notations from Campbell & Stanley (1963):

O = Data handling assessment

R = Randomisation at pupil level

X_1 = Group size 1:1 X_2 = Group size 1:2 X_3 = Group size 1:4

R	O	X_1	O
R	O	X_2	O
R	O	X_3	O

4.3.3 Participants

The study included a total of 176 pupils from 11 different schools, with 67 peer tutors and 109 tutees. A detailed analysis of the baseline characteristics of the tutees, tutors and schools can be found in Chapter 5 for the analysis of the results.

Prior to randomisation, each participating secondary school identified 7 pupils and each primary school identified 12 pupils who were eligible to participate in the trial. Since the removal of the national framework of 'levels' to report pupils attainment and progress it was not possible to specify specific inclusion criteria. A relative criterion, depending on each individual school assessment – "above expected", "expected" and "below expected" progress will be used. Consequently, consistency between individual schools were a potential limitation.

The inclusion criteria used in the study are:

School inclusion criteria: Secondary and primary schools are eligible to take part in the trial if they agree to all trial procedures including provision of pupil data, informing parents, randomisation and implementation of the intervention.

Primary pupil inclusion criteria: Pupils eligible for the intervention will be either at "expected" or "above expected" for progress in Mathematics by the end of Year 6, based on teacher assessments.

Secondary pupil inclusion criteria: Pupils eligible for the intervention will be either at "expected" or "above expected" for progress in Mathematics by the end of Year 8, based on teacher assessments.

4.3.4 Randomisation

Randomisation was performed at an individual level in each of the participating schools and carried out by an independent person[5]. The reason randomisation is blinded is to prevent selection bias, as a recent systematic review found that trials using non-blinded assessors generated more optimistic effect sizes than blinded assessors (Hróbjartsson et al., 2014).

Unequal randomisation was used for the peer tutors in each secondary school participating in the study, with 1 pupil assigned to the 1:4 group, 2 pupils assigned to the 1:2 group and 4 pupils assigned to the 1:1 group. The learners were not aware at the start of the trial that they had been assigned to different sized groups. The purpose of using unequal randomisation for the online peer tutors means that the only difference between the two groups is the group size as all other characteristics are distributed evenly across the two groups (Torgerson & Torgerson, 2001). However, after the second cohort of schools had started a limitation and potential source of bias was identified. As the randomisation was not sequence generated, such as random-number table or a computerised random number generator, the method used has the potential to introduce unintentional bias into the study. Peer tutors were assigned to groups through names being placed onto a folded piece of paper and drawn by independent person to be assigned to a group.

The decision was taken to continue to use this method of randomisation and acknowledge the limitation in the discussion.

[5] A qualified teacher not involved in the study performed the randomisation, after performing a trial run under the supervision of the author of this study to ensure the process was completed correctly.

For each participating primary school, the pupils were ranked in order of performance based on the data handling pre-assessment as a measure of mathematics ability for this specific topic, with 1 equating to the highest and 12 the lowest scores These were divided into groups of three, starting with 1, 2, and 3. These were randomly selected by the independent person to be allocated into the group sizes 1:1, 1:2 and 1:4. This process was repeated for blocks 4, 5, 6 and 7, 8, 9, then finally 10, 11, 12. Through blocking the performance ability as indicated by the pre-assessment on data handling this would result in an equal distribution of ability across the groups.

4.3.5 Outcome measures

The assessment of the primary outcome used was designed by the researcher in partnership with a Year 6 teacher from school B, using previous standardised SAT assessments from the schools TestBase assessment system. In the previous research a standardised assessment (Centre for Evaluation and Monitoring InCAS) was used as the use of researcher made assessments tend to be associated with larger effect sizes (Cheung & Slavin, 2016) and have the potential to introduce bias into the study. However, due to financial restrictions the potential cost of using InCAS would be in excess of £3000 and not a viable option for this thesis study.

The assessment was piloted with 79 pupils in cohort one and the reliability of the assessment checked using Cronbach's Alpha and person and item reliability using Rasch measurement to ensure that this was a reliable and valid measure for the trial. The analysis and discussion of these results are included in the data management section of this chapter.

Primary outcome

The primary outcome was the data handling score on the assessment designed using previous standardised SAT questions. The assessment consisted of a total of 17 marks and involved interpreting line graphs, interpreting pie charts and analysing scatter graphs. In the assessment, a higher score represented higher attainment.

Administration of outcome measure

Teachers were asked to administer the outcome measure on an agreed date prior to the start of the intervention and in the final week of the planned intervention. It was agreed that the researcher would

be present and they were asked to administer the test under 'exam' conditions with all pupils who participated in the research at the same time.

4.3.6 Sample size

The power calculations are based upon the following. As the intervention is comparing the group sizes 1:2 and 1:4 with the active control 1:1, it is reasonable to conclude that the differences are likely to be smaller than the difference between a standard business as a usual control[6].

It is important to note that the pilot study will not be sufficiently powered from the three cohorts of schools in this study (Kraemer & Blasey, 2016). The following power calculation provides a theoretical sample size for the aggregated trials to reach a sufficient sample size to allow generalisations to be inferred (and to balance the risks of Type 1 and Type 2 errors).

The power calculations are based on the following assumptions:

1) A school-level intra-cluster correlation (ICC) of 0.15
2) Equal cluster sizes of 4 per school (3 groups of 4 per school)
3) 30 Schools (clusters) participating in the aggregated trial
4) 80% power for 95% confidence interval

For the purposes of calculating the sample size it is assumed that 30 schools (online tuition groups) would be recruited over the duration of the aggregated trial with 12 pupils per primary school; this would result in a total sample size of 360 pupils. Assuming 4 pupils per school were allocated to the 1:1 group and an intra-cluster correlation coefficient of 0.15, based upon the assumptions, the trial would be able to detect an effect of 0.19 standard deviations with 80% certainty (Kraemer & Blasey, 2016).

4.3.7 Recruitment

The schools targeted for the study were strategically recruited from the North East of England due to the logistical demands of the efficacy study, with the exception of one school partnership in the

[6] Group A (1:1) is the active control and the comparisons are B (1:2) vs A and C (1:4) vs A. Therefore it is not required to take into account any increased type 1 error.

midlands who were recruited through the online platform learning provider. A presentation was delivered at a Durham LEA conference on the previous feasibility study for the online cross-age peer tuition trial in September 2015 and five secondary schools expressed an interest participating in the research. Secondary school A recruited four feeder primary schools (schools B, C, H and I), secondary school D recruited primary school E, secondary school F recruited primary school G and secondary school J recruited primary school K. Unfortunately a school partnership L and M were unable to agree a suitable time for the delivery of the intervention so these withdrew from the study before selecting students and signing an agreement to participate in the study.

Schools who wanted to take part were asked to sign an 'Agreement to Participate Form' (Appendix 2) to ensure they agreed to all trial-related procedures. All schools informed parents of the study using material provided (Appendix 2) with the opportunity to withdraw their child's data from being used in the evaluation (opt out) prior to randomisation. Only two parents withdrew consent prior to the pre-assessment and randomisation process, these pupils were replaced by the school. Participating schools then shared pupil data (including pupil name, date of birth (DOB) and pupil premium status).

4.4 Data management

4.4.1 Data collection methods

The trial protocol and Statistical Analysis Plan (SAP) are included in Appendix 3.

Study instruments

The primary assessment used in this study was created by the researcher in partnership with a Year 6 teacher in school B using previous standardised assessment material available in the school. As previously explained, the assessment was piloted with 76 pupils in cohort one to enable the Cronbach's Alpha reliability and RASCH analysis to be completed.

The Cronbach's Alpha

The assessment included a total of 17 items and using SPSS software, the Cronbach's Alpha coefficient was calculated. (Gliem & Gliem, 2003) state that the closer the Cronbach's alpha coefficient is to 1.0 the higher the internal consistency of the items.

The analysis was completed three times. Firstly, the Cronbach's alpha was calculated for the tutors (N=31), this was 0.72 Secondly, the Cronbach's alpha was calculated for the tutees (N = 48), this was 0.734. Finally, as both the tutors and tutees completed the same assessment these were combined

(N= 79) for a Cronbach's alpha 0.916. This demonstrated that the assessment had a high internal consistency for the items.

RASCH Analysis

The RASCH analysis was completed using the WINSTEPS software version (3.81.0). Firstly the data was analysed for the item map for the tutees then tutors, then combined for the total sample of 79 students.

The RASCH analysis for the combined tutor and tutees reported a person reliability 0.86 and item reliability 0.96.

The analysis of the peer tutors only reported a person reliability 0.70 and item reliability 0.85. Replicating the analysis for the tutees only reported a person reliability 0.65 and item reliability 0.86.

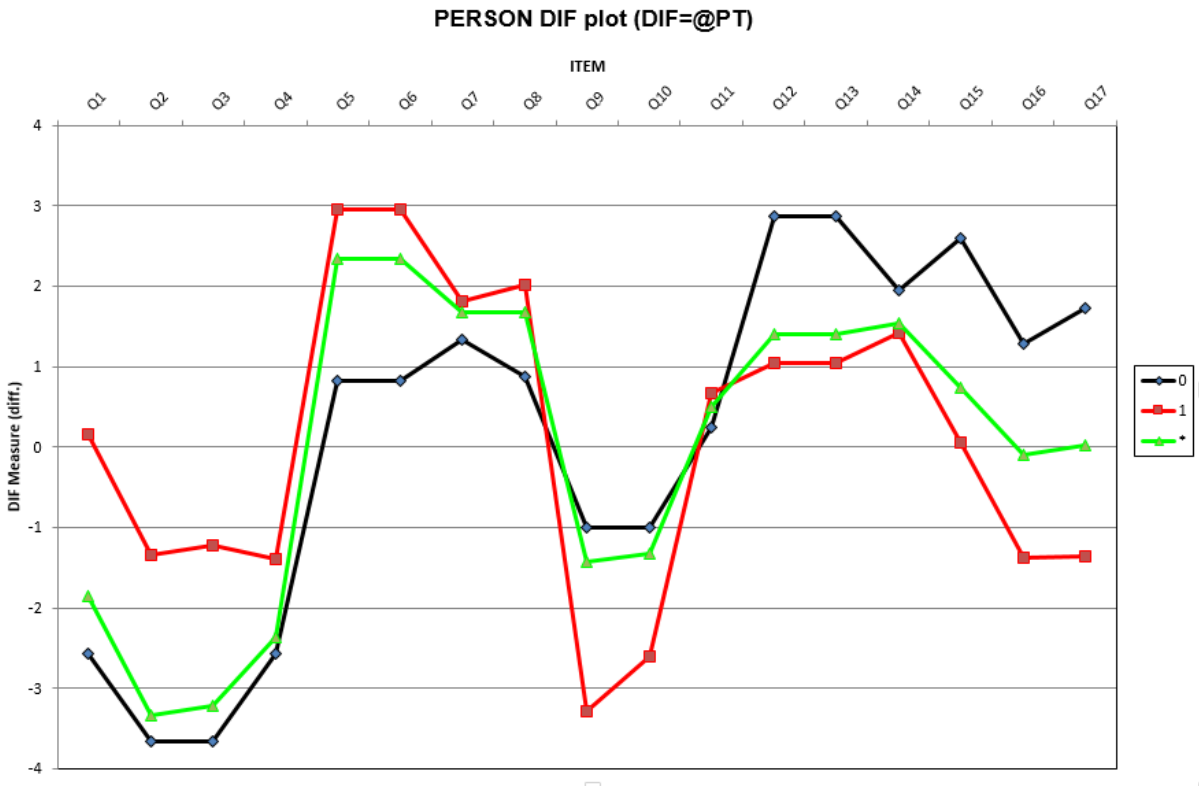


Figure 14: A plot of the DIF measure against the person DIF plot.

Figure 14 reports the difficulty of the item for each person classification and shows a strong correlation between the RASCH scaling item. The graph shows the tutees (0), peer tutors (1) and combined tutees and tutors (*). As will be explained in Figure 15, a flip occurs in the difficulty between the tutees and tutors with this issue likely due to the design of question 1 rather than due to ability.

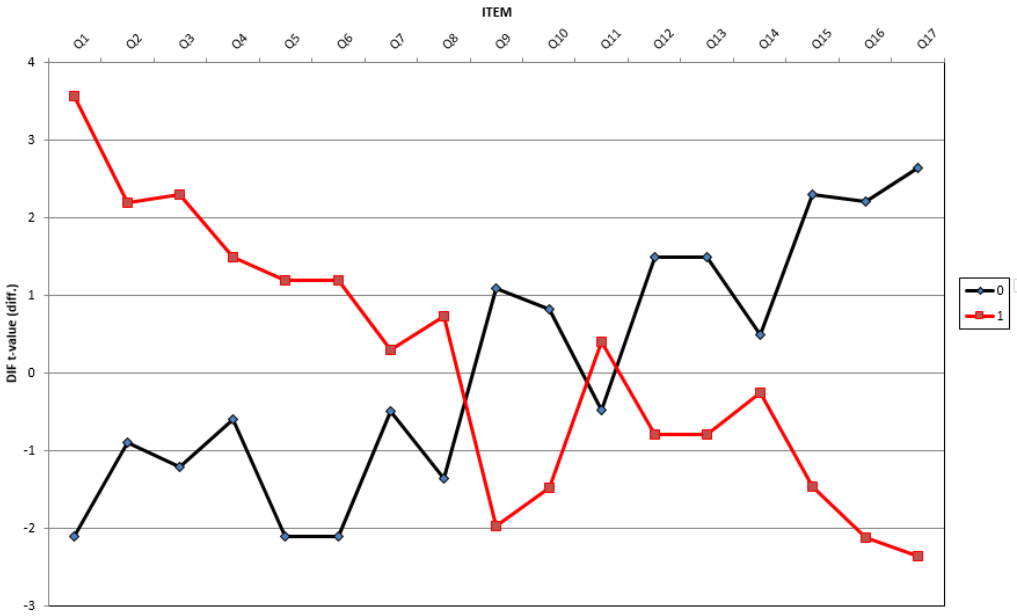


Figure 15: DIF t-test of the item DIF against the overall difficulty.

Figure 15 is interesting as question 1 indicates a difference which is greater than 2 ± between the tutors and tutees. Question 1 represented the easiest challenge question and it required pupils to interpret data from a line graph, with a very specific answer of 7.(00) provided in the SAT mark scheme. A number of older peer tutors provided answers just as 7.05 which is incorrect according to the mark scheme. This did not appear to significantly impact the reliability of the assessment and due to restrictions with the study (unable to repeat a pilot study) it was decided to keep this question.
Question 17 presented the most challenging question relating to difficulty so it is not surprising to see a difference between the tutors and tutees.

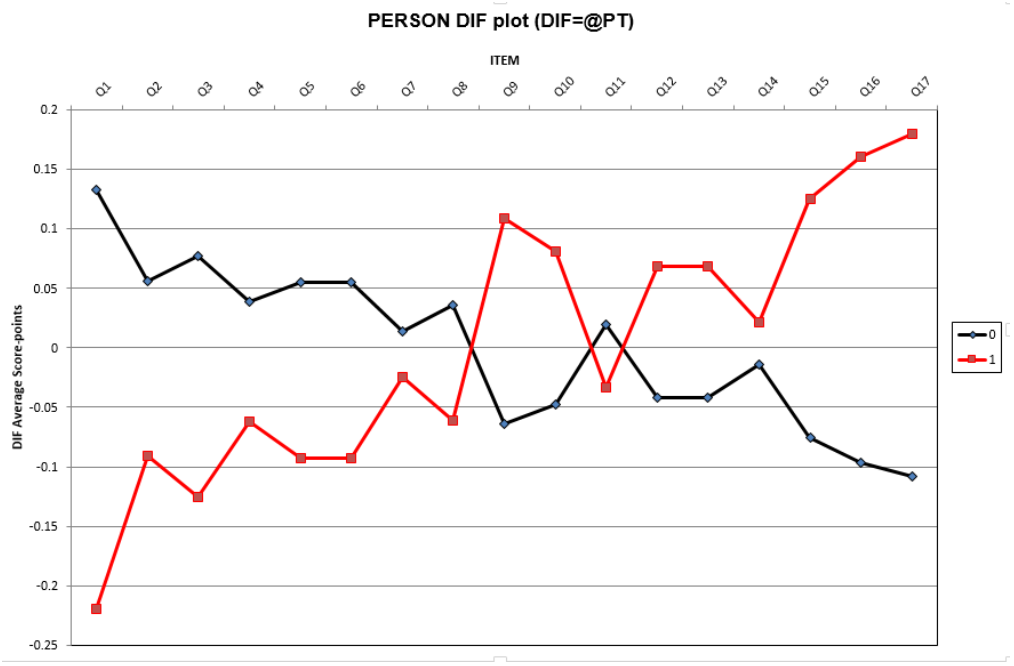


Figure 16: The average difference between the observed response and the expected outcome for each person CLASS on each item.

Figure 16 shows potential bias in questions 1, 16 and 17 due to the difference between tutees (0) and tutors (1). However, with the RASCH analysis for the assessment reporting an overall person reliability of 0.86 and an item reliability of 0.96, a pragmatic decision was made to use this as an alternative to the standardised InCAS assessments previously used.

4.4.2 Trial diagram for pilot study (cohort 1)

The trial diagram outlines the pilot study for cohort one. A consort flow diagram for the full aggregated trial is included in the impact evaluation of data in Chapter 5 of this thesis.

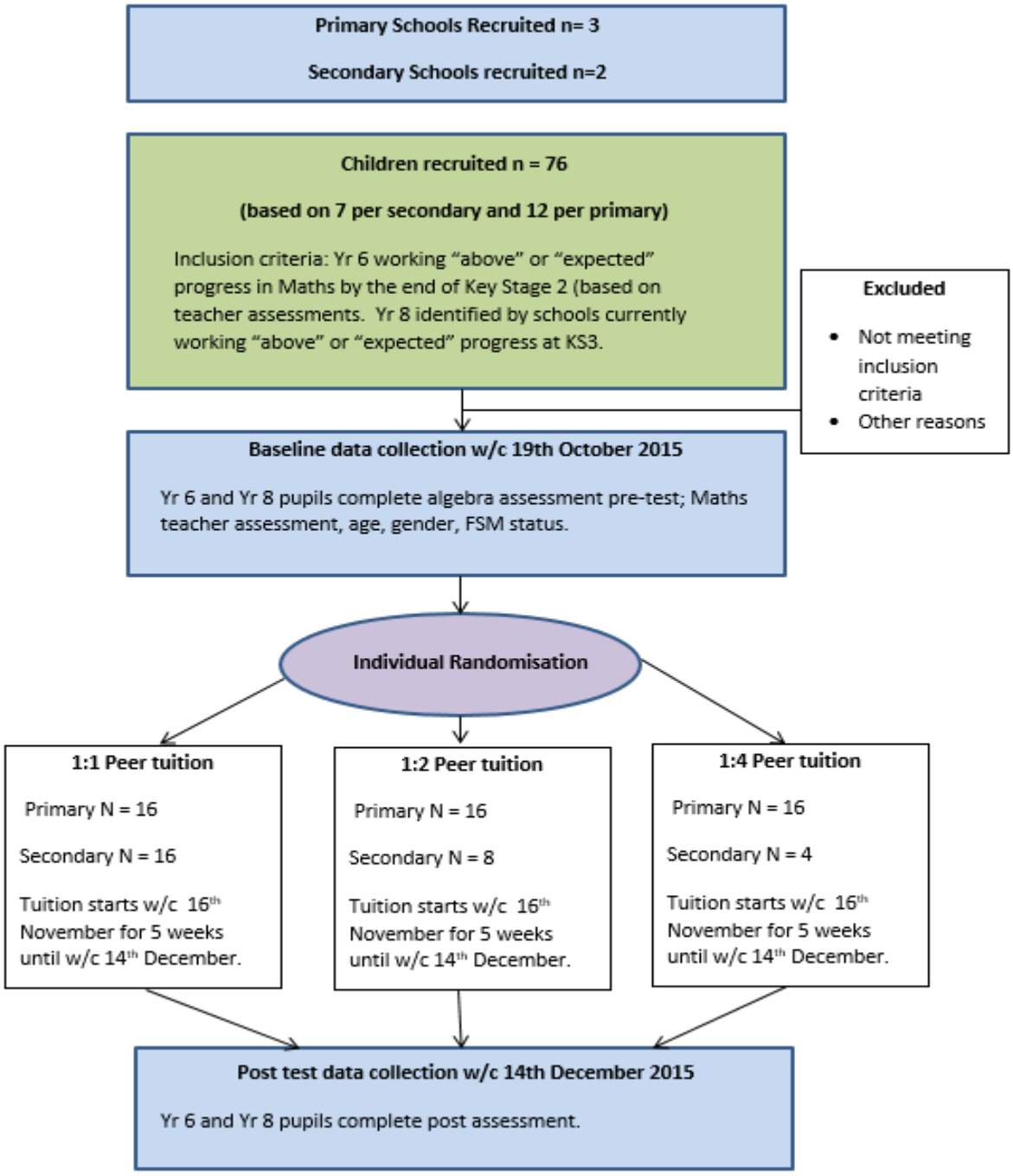


Figure 17: Trial diagram Study One cohort 1

4.4.3 Data Retention

The primary plan for the retention of participants involved regular weekly contact with each of the participating schools due to the planned nature of efficacy study. The small-scale nature of the trials ensured that it was feasible for the researcher to observe and closely monitor the intervention. In addition, this enabled a close working partnership with the schools as the researcher would often be required to trouble shoot IT issues at the primary school during the first two weeks of the intervention delivery.

If a pupil missed the assessment through absence, teachers ensured that the pupils completed this task on the day they returned to school with the assessment then collected from the school. Through working closely with schools this relationship was possible to establish. In the instance of a pupil leaving the school, it would not be possible to complete the assessment.

4.4.4 Statistical methods

The statistical analysis plan is available in Appendix 3. The analysis was conducted using the R software (Version 3.3.1) and EEF Analytics package (Kasim, ZhiMin & Higgins, 2015). Impacts were estimated on the basis of intention to treat, whereby all pupils who were initially involved in the pre – and post-testing were analysed according to the group they were initially assigned, regardless of whether they went on to participate in the intervention.

Effect sizes were calculated and are presented alongside their 95% confidence intervals. The effect size reported in this study use Hedges' g instead of the traditional Cohen's d. The reasoning for this is that both Cohen's d and Hedges' g pool variances on the assumption of equal population variance, however g pools using n-1 for each sample instead of n, providing a better estimate with smaller sample sizes (Grissom & Kim, 2005).

4.4.4.1 Primary analysis

A linear mixed effects model fit by REML (Restricted Maximum Likelihood) was used for the primary analysis. The model used pre- and post- assessments as fixed and schools as random effects. The justification for schools as random is because it is assumed the effect varies randomly within the population of the organisation. The reasoning for selecting pre- and post-assessments as fixed is because it is assumed to be measured without measurement error and the variable used contains

most or all of the variable values in the population. The underlying assumptions will be discussed in detail in the analysis of the data in Chapter 5.

4.4.4.2 Cumulative meta-analysis

The data collected from the three cohorts of trials are aggregated using a cumulative meta-analysis and presented using a box plot graph.

4.5 Implementation and process evaluations

The process evaluation will have three purposes. Firstly, it will assess the fidelity of the delivery of the intervention, as this is fundamental prerequisite for the intervention. Secondly, it addressed the feasibility of the intervention for an aggregated trials methodology. Thirdly, it will support the impact evaluation to determine if the group sizes 1:2 and 1:4 can be as effective as 1:1 online peer tuition.

4.5.1 Process Evaluation Research questions

The process evaluation questions are:

- How feasible is it to implement an online cross-age peer tuition project across the transition boundary of primary to secondary school as an aggregated trial?
- How feasible and acceptable do teachers feel it is for using online peer tuition as a transition intervention?
- What are the barriers to the successful implementation of online cross-age peer tuition as a transition intervention? How can these be minimised in future aggregated trials?
- What are the views of the intervention, of the peer tutors, tutees, teachers and IT technicians?
- What are the staff perceptions of the current and possibly sustained impact of the intervention on children's educational attainment?
- What are their suggestions for change if the intervention was to be more widely implemented?

4.5.2 Methods

In the planning stage of the process evaluation, a number of methodological considerations were made using a pragmatic approach to data collection. The process evaluation consisted of three main sections: set-up, lesson observations and discussions from participants after the completion of the trial.

Firstly, due to the potential technical issues in schools the intervention required the researcher to be attend the online lessons to supervise the set up and access to the online platform. Therefore the first methodological consideration involved the use of lesson observations to monitor the fidelity of the intervention. Ciesielska, Bostrum & Ohlander, (2018) identify three positions a researcher can assume during observation: completely, partially or non-participating. The role of the observer in this study was as a partial participant, due to the requirement to intervene for IT issues and to support peer tutors log into the virtual classroom. The observations were completed in the secondary schools as this is where the majority of technical issues were encountered during the previous online cross-age peer tuition trials (Harrison, 2015). A fidelity measure was recorded on a scale of 0 to 9 for each lesson observed. Alongside the fidelity measure, note-taking was used to record issues or observations from the lesson.

Secondly, consideration was required for discussing how the participants had engaged with the intervention. The decision to use focus groups for the learners and peer tutors was based on the two considerations. As the learners and peer tutors were children, the use of focus group interviews would help establish a safe environment. Morgan (1996) suggests the use of focus groups create safe environments for the informants and allow the researcher to collect a group accumulation of statements, opinions and experiences. The focus groups were semi-structured, as Mason (1994) highlights the importance of providing a lightly structured conversation for interventions which are relatively short. The exception would be when conducting long term field work as unstructured approaches would allow respondents to express their thought sin their own time and pace. As the intervention was delivered over a short timescale, a semi-structured approach was used. The second consideration was time available to speak with the learners and peer tutors. A focus groups approach rather than interviews allowed for a shorter period of time and fit around lessons in schools.

Finally, interviews were conducted with teachers rather than a focus group approach. The justification was logistical, as access to teachers in their school setting would be easier than asking teachers to travel to a central location at a specific point in time. Again, a semi-structured approach was used for the format due to potential issues with unstructured interviews (Corbin & Morse, 2003).

4.5.3 Fidelity

Lesson observations were completed for all the online peer tuition sessions delivered over the three cohorts in the aggregated trials. Fidelity was assessed for every intervention group in the trial using a measure created for the trial. The measure consisted of three component scores relating to (a) how well the technology worked (b) the tutor attitudes to learning (c) content delivered in the 30 minutes. For each of these components was rated between 1 and 3, with 1 corresponding to 'poor', 2 corresponding to 'worked partially' and 3 corresponding to 'worked as planned'. Therefore, each online tuition session for a school partnership had a fidelity score which ranged from 3 to 9. Lower scores corresponded with lower fidelity and individual lesson comments were recorded as part of the process evaluation. Appendix 12 provides a summary of the fidelity measures.

4.5.4 Ethics and dissemination

4.5.4.1 Ethical approval

Approval from Durham University Ethics Committee was sought before commencing the project[7] and granted on 25th September 2015. Additional ethics amendments were accepted on 19th January 2016 for the cohort 2 and 29th April 2016 for cohort 3.

Schools informed parents of the pupils participating in the study with a parental information and opt-out letter. The opt-out letter allowed parents the opportunity to withdraw their child's data at any point in the study.

As previously mentioned, the design used an active control through the use of the 1:1 group and no pupils were withheld the intervention. In addition to all the pupils receiving the intervention at the same time, the intervention involved the use of standard learning material and content already in use in schools.

4.5.4.2 Protocol amendments

Prior to starting the pilot an overview of the trial was submitted with the ethics application to Durham University Ethics Committee. This is the version 1 of the very early stage protocol and a copy is attached in Appendix 1.

[7] Appendix 4 includes the ethics proposal, parental letter, opt-out form and ethics approval letter.

Once the pilot had started and new knowledge on the importance of protocols in trials was acquired as part of the literature review for chapter 2, an updated protocol was written. This version can be found in Appendix 2 and this is acknowledged as a limitation in the discussion.

Amendments to version 2 are:

1. Initial trial outline updated based on research evidence from protocols.
2. School D and E unable to complete the online peer tuition due to technical issues with the primary school firewall (discussed in process evaluation), therefore these created a small retrospective control as they agreed to compete the assessments pre and post.
3. School K requested an additional participant therefore an extra one to one group is included, pragmatic decision to increase the sample size based on a school request.
4. Error in the initial power calculation corrected, as the sample size per school reduced from 12 to 4 (1:1 group size) as a conservative measure.

4.5.4.3 Confidentiality

As part of the previous trial for the feasibility of online cross-age peer tuition (Harrison, 2015), a data protection training course was completed prior to the collection of data. The data is stored on a password protected computer and not on unsecure hard drives, backed up on the university server. All student names and school identifications are replaced with anonymous codes to ensure no identification is possible.

4.5.4.4 Dissemination

As part of the dissemination of the research presentations were given at the Durham LEA pupil premium conference in September 2016 and More Able and Talented Conference in January 2017. At the end of each cohort, a report is generated for the participating schools with individual data for test scores for pupils from their school. This allowed schools to evidence the impact of the intervention for Ofsted or LEA inspections. The findings of the aggregated trial will be written up and combined with the theoretical framework for school based aggregated trials linked to prospective cumulative meta-analysis for peer review and publication.

4.6 Methodology Summary – Study 1

The first study in this thesis uses a randomised controlled trial design to test the effectiveness of online peer tuition group sizes for 1:1, 1:2 and 1:4. Through the use of an aggregated trials approach and the replication of the pilot study over a sequential time period, the study will provide empirical data to support the theoretical framework for using small scale aggregated trials linked to a prospective cumulative meta-analysis.

Table 5 provides an overview of the research questions involved in study 1, with a description of the method, data collected and analysis. The full analysis for the impact and process evaluation for study 1 can be found in Chapter 6.

Table 5: Overview of study 1

Research question	Method	Data	Analysis
Is the implementation of an online cross-age peer tuition intervention feasible between primary and secondary schools as an aggregated trial?	Interviews, focus groups and lesson observations	Transcriptions of interviews and focus groups. Lesson observations.	Analysed using an inductive approach to Thematic Analysis (NVIVO software)
Is online cross-age peer tuition a feasible intervention strategy between schools as part of an aggregated trial?	Interviews, focus groups and lesson observations	Transcriptions of interviews and focus groups. Lesson observations.	Analysed using an inductive approach to Thematic Analysis (NVIVO software)
Are the instruments used appropriate for the ability of students involved in the study?	Pre- assessment	Pre-assessment data	Cronbach Alpha and RASCH
Can the intervention maintain fidelity if delivered to schools in other locations in the UK?	Process evaluation (observation, interviews and focus groups)	Lesson observation and fidelity score	Comparison of the fidelity scores and lesson observation analysis
Does one-to-two online cross-age peer-tutoring produce equivalent effects in mathematics attainment[8] compared with online one-to-one tuition?	Randomised Controlled Trial	Pre-test, post-test, dosage of intervention.	Multi-level model (linear mixed effects fit by REML)
Does one-to-many online cross-age peer-tutoring produce equivalent effects in mathematics attainment[2] as compared with one-to-one tuition?	Randomised Controlled Trial	Pre-test, post-test, dosage of intervention.	Multi-level model (linear mixed effects fit by REML)

[8] Mathematics attainment is defined as mastery level concepts in KS2 curriculum for Algebra.

Chapter 5: Impact and process evaluation for study 1 into the effectiveness of online peer tuition group sizes.

The focus of this chapter is to present the findings from Study One, outlined in Chapter 4, investigating the feasibility and relative effectiveness of online cross-age peer tuition with differing group sizes across the transition boundary between primary and secondary schools. The first section outlines the impact evaluation and the second section presents the process evaluation. The initial design of the aggregated RCT planned three data collection periods to evaluate the effectiveness of online peer tutoring for the group sizes 1:1, 1:2 and 1:4 on the mathematics attainment of Year 6 pupils. The pilot was conducted as an efficacy study with the sample size limited due to resource constraints. Therefore, it is important to note that the conclusions from this study are not intended to be generalizable.

The intervention was designed to allow Year 8 secondary pupils (12-13 years) to peer tutor Year 6 pupils (10 – 11 years) on the topic of data handling in mathematics. The topic was chosen through discussions with the initial feeder primary schools identified for the pilot study, as this is often an area of weakness in KS2 mathematics. The lessons are delivered online by the Year 8 peer tutors for 30 minutes for 5 weeks, using the TUTE online learning platform.

Prior to the start of the intervention, the dates and times were co-ordinated and agreed between the participating primary and secondary schools. The aggregated trial includes three cohorts: cohort 1 delivered the tuition between November – December 2015; cohort 2 delivered the tuition between February – March 2016; and cohort 3 delivered the tuition between June – July 2016.

In advance of the presentation of the study, it is useful to revisit the primary and secondary research questions. Table 6 provides an overview of the research questions, method, data and analysis used in the pilot efficacy study and first three cohorts of the small scale randomised controlled trials for the development of a prospective cumulative meta-analysis.

Table 6: Overview of Study 1 research questions, methods, data and analysis.

Research question	Method	Data	Analysis
Is the implementation of an online cross-age peer tuition intervention feasible between primary and secondary schools as an aggregated trial?	Interviews, focus groups and lesson observations	Transcriptions of interviews and focus groups. Lesson observations.	Analysed using an inductive approach to Thematic Analysis (NVIVO software)
Is online cross-age peer tuition a feasible intervention strategy between schools as part of an aggregated trial?	Interviews, focus groups and lesson observations	Transcriptions of interviews and focus groups. Lesson observations.	Analysed using an inductive approach to Thematic Analysis (NVIVO software)
Are the instruments used appropriate for the ability of students involved in the study?	Pre- assessment	Pre-assessment data	Cronbach's alpha and Rasch analysis
Can the intervention maintain fidelity if delivered to schools in other locations in the UK?	Process evaluation (observation, interviews and focus groups)	Lesson observation and fidelity score	Comparison of the fidelity scores and lesson observation analysis
Does one-to-two online cross-age peer-tutoring produce equivalent effects in mathematics attainment[9] compared with online one-to-one tuition?	Randomised Controlled Trial	Pre-test, post-test, dosage of intervention.	Multi-level model (linear mixed effects fit by REML)
Does one-to-many online cross-age peer-tutoring produce equivalent effects in mathematics attainment[2] as compared with one-to-one tuition?	Randomised Controlled Trial	Pre-test, post-test, dosage of intervention.	Multi-level model (linear mixed effects fit by REML)

[9] Mathematics attainment is defined as mastery level of concepts in KS2 curriculum for Algebra.

5.1 Logic model / theory of change

A logic model or theory of change (Table 7) reflects the necessary conditions (inputs and processes in the school) for their intervention to be successful. The primary purpose of the logic model is to provide the underlying assumptions about how the intervention will enable the expected outcomes to be achieved (Clapham et al., 2017). The process evaluation will provide evidence on whether these conditions were successfully achieved in the implementation of the intervention[10].

Table 7: Logic model

Inputs	Processes in school	Change mechanism	Pupil and mentor intermediate impacts	Outcomes
• Sufficient computers with access to the internet and headsets with microphones • Good classroom environment sufficient space between students on the computers • Online peer tuition resources to scaffold the learning • In school supervision (TA or teacher)	• Online 30 minute peer tuition lessons delivered over 5 weeks • Online peer tuition lessons in mathematics for the topic data handling • TA or teacher to supervise log in and behaviour during the lessons	**Education** • Children learn mathematics with a peer tutor in an online synchronous learning environment • Peer tutors use engaging resources to scaffold learning • Additional learning support to complement classroom practice **Online delivery** • Online delivery to engage students learn mathematics via a different medium to traditional classroom learning • Varying group sizes allows more opportunities for students to ask questions and receive support from peers	• Improved concentration and engagement in mathematics • Improved confidence and self-esteem • Improved mathematical reasoning skills • Improved communication skills	• Increase in mathematics attainment on data handling assessment

[10] The intervention involves online peer tutoring for the group sizes 1:1, 1:2 and 1:4 in mathematics

5.2 Impact evaluation

5.2.1 Participants

Schools were allocated to a cohort for the delivery of the intervention in either the autumn term 2 (November – December 2015), spring term 2 (March – April 2016) or summer term 2 (June – July 2016).

The three cohorts of schools included a total of 176 pupils from 11 schools, 67 peer tutors and 109 tutees. Each participating secondary school in the trial was asked to identify 7 pupils as peer tutors, with 4 pupils randomly (unequal randomisation) assigned to 1:1 grouping, 2 pupils into 1:2 grouping and 1 pupil into the 1:4 group. In each primary school, the pupils were ranked in order of performance based on the data handling pre-assessment (1 equating to best and 12 the least). These were divided into groups of three, starting with 1, 2 and 3. These were then randomly selected by an independent person to be allocated into the group of 1:1, 1:2 and 1:4. This process was repeated for the blocks 4,5,6 and 7,8,9 and 10, 11, 12. Blocking performance ability as indicated by the data handling pre-assessment aimed to produce an even distribution of attainment across the groups.

The method for randomisation replicated the previous research for the MA feasibility study, whereby tutors names were folded onto paper and drawn by an independent person (teacher) to allocate to the group sizes 1:1, 1:2 and 1:4. This is acknowledged as a limitation within the study, as when I learned about the importance of the protocols structure the importance of randomisation recording, such as computer randomisation software became apparent. However, as cohort 1 had already started the decision was taken to continue with this approach for the next two cohorts.

<u>Cohort 1</u>

As the CONSORT diagram in Figure 18 shows, five schools agreed to participate in the trial in cohort 1 with a sample size of 79 pupils. Prior to randomisation, two schools were unable to continue with

the delivery of the intervention due to technical issues with the firewall configuration in the primary school. Consequently, 21 pupils were excluded after the initial pre-tests had been completed. The school agreed to act as a respective control and completed the post assessment as planned. The data is not included in the main analysis, however the analysis data is provided in Appendix 5. Furthermore, a primary pupil withdrew from the programme prior to the pre-test as the pupil did not wish to participate. During the data collection stage, two pupils were unable to complete the post-assessment due to school absences.

Therefore, 56 pupils completed the online peer tuition in cohort one with 1:1 (n=26), 1:2 (n=18) and 1:4 (n=14).

Attrition in this cohort with the inclusion of the two schools who were unable to participate was 29.2%. In the analysis of the primary data, attrition for cohort 1 was only 3.5% with missing data for only one pupil in the 1:2 and 1:4 groups.

Cohort 2

As the CONSORT diagram in figure 19 shows, two schools agreed to participate in the trial in cohort 2 with a sample size of 38 pupils. During the data collection stage, one pupil was unable to complete the post-assessment due to school absence. Therefore, 38 pupils completed the online peer tuition in cohort one with 1:1 (n=16), 1:2 (n=12) and 1:4 (n=14).

In the analysis of the primary data, attrition for cohort 2 was 2.6% with missing data for only one pupil in the 1:4 group.

Cohort 3

As the CONSORT diagram in Figure 20 shows, five schools agreed to participate in the trial in cohort 3 with a sample size of 59 pupils. All pupils completed the intervention (n=59) with group sizes 1:1

(n=26), 1:2 (n=18) and 1:4 (n=15). One pupil could not complete the post-assessment due to medical issues.

In the analysis of the primary data, attrition for cohort 3 was 1.7% with missing data for only one pupil in the 1:1 group.

5.2.2 CONSORT Flow Diagram

Figure 18: CONSORT flow diagram cohort 1

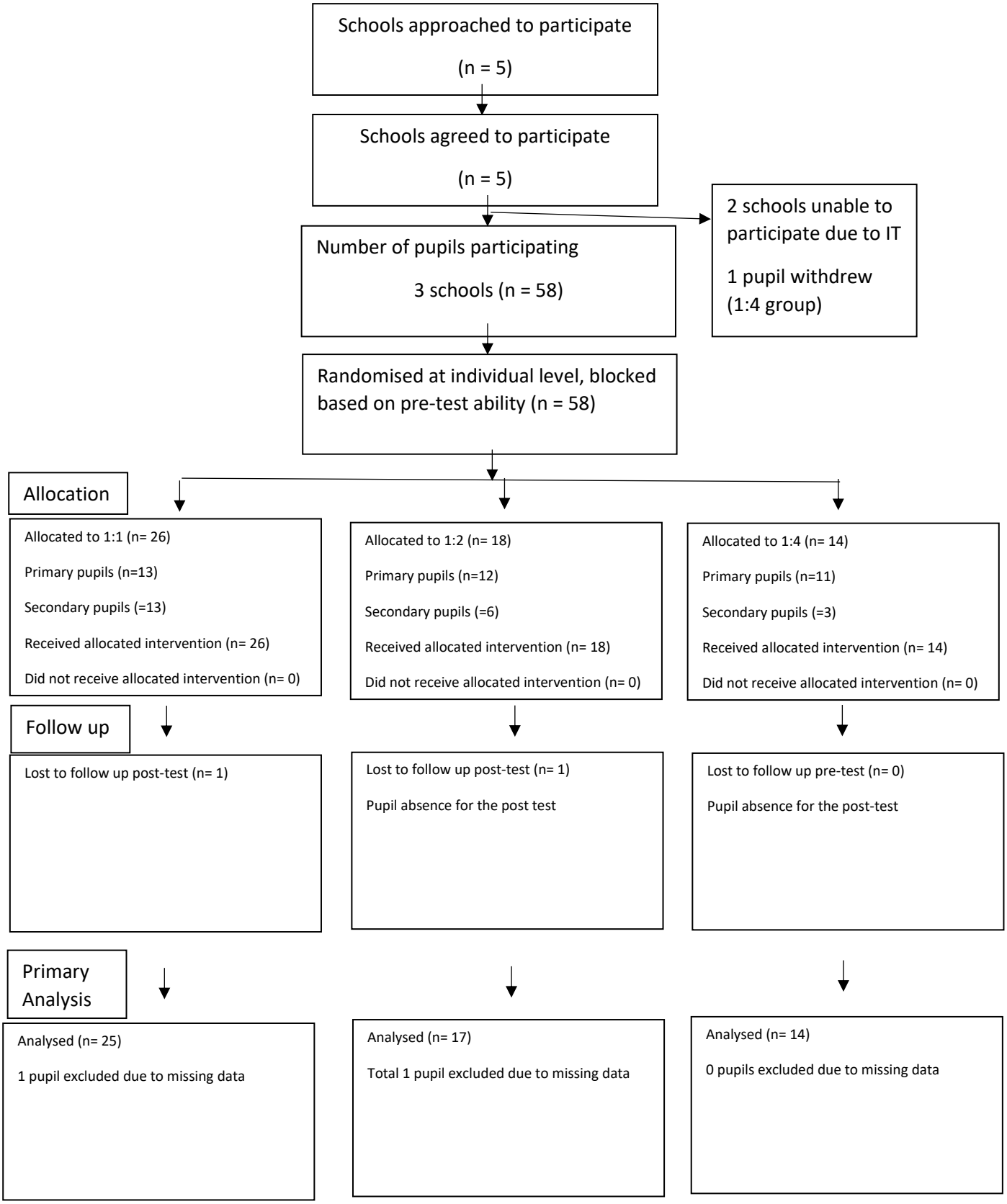


Figure 19: CONSORT flow diagram cohort 2

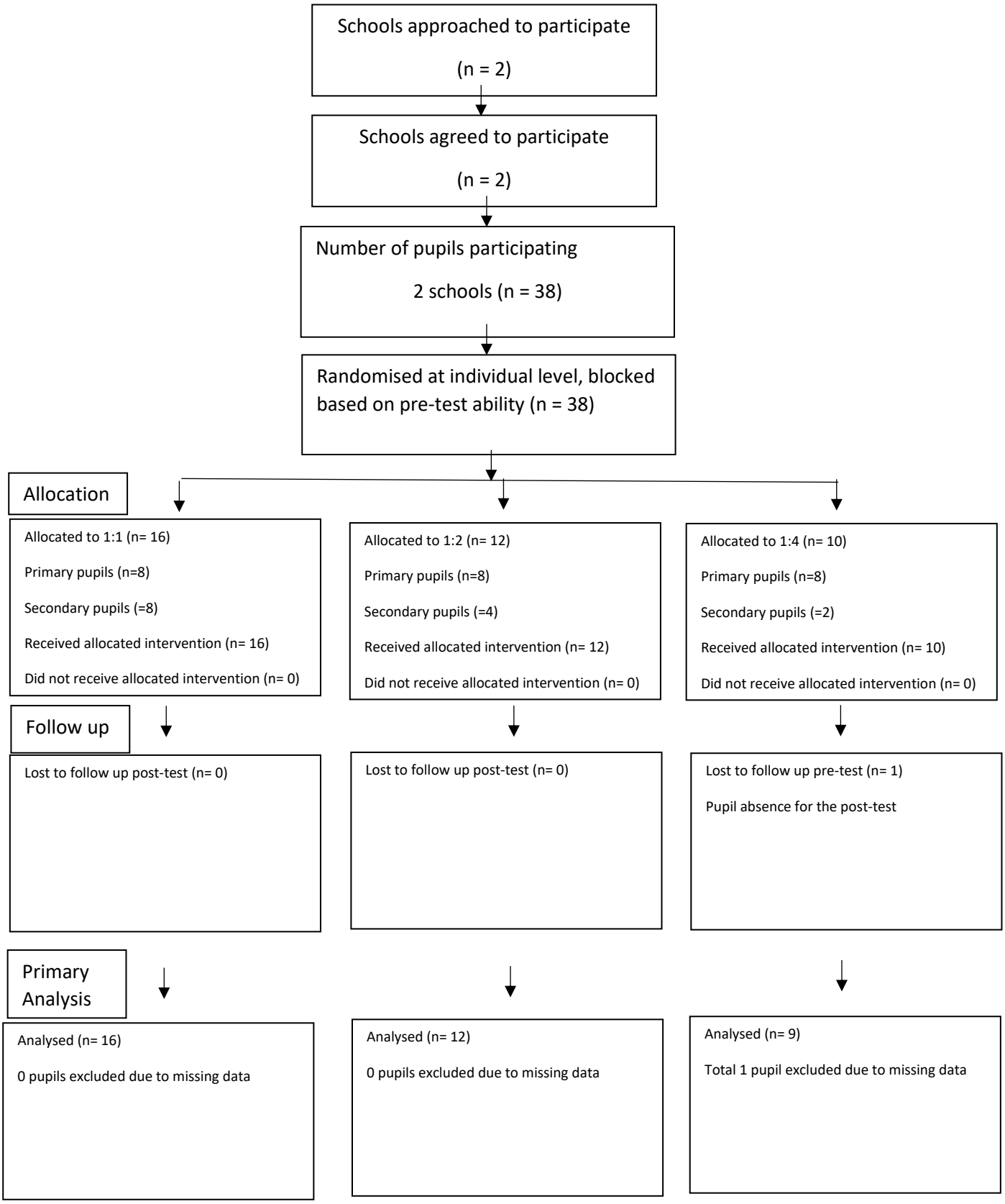
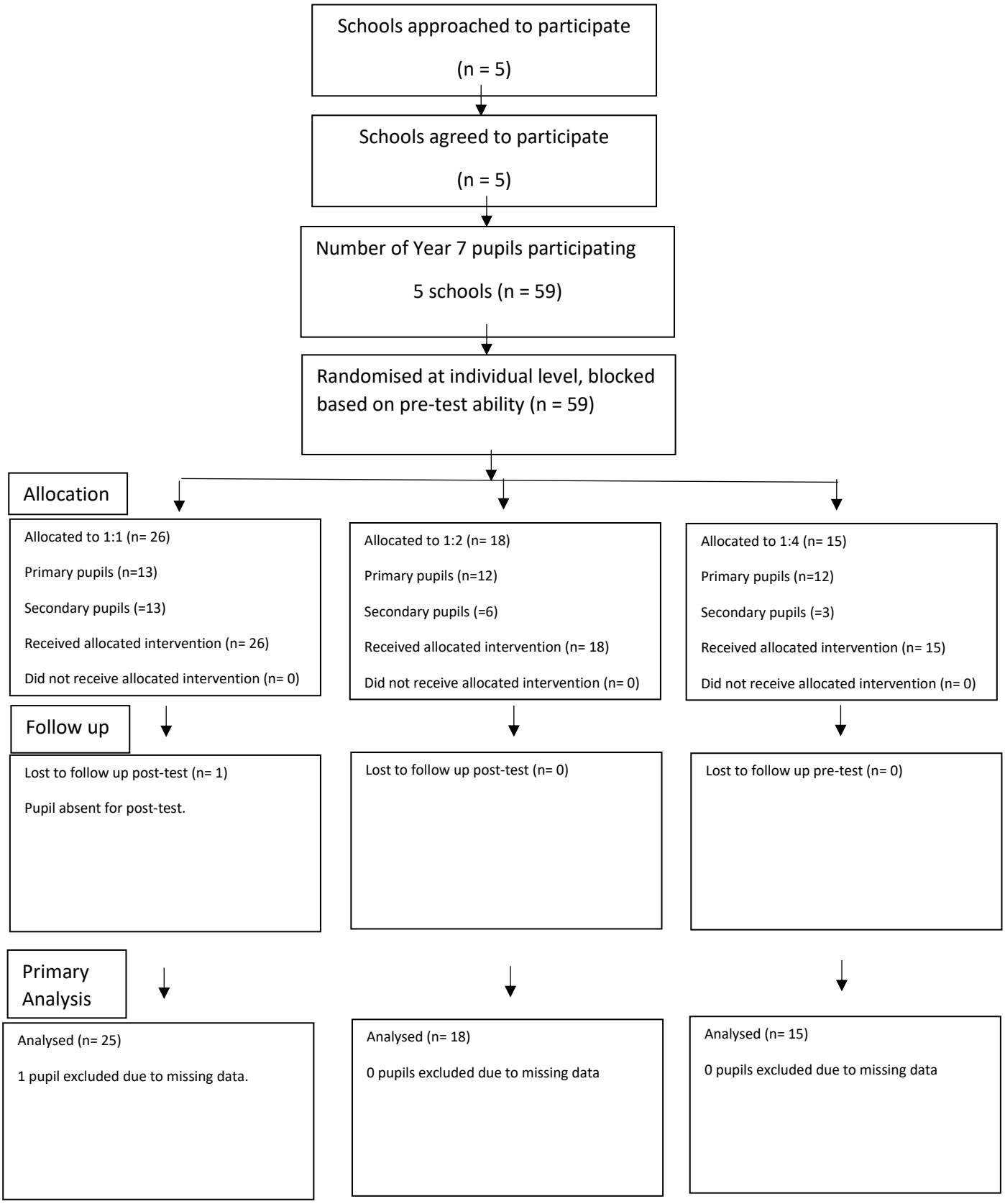


Figure 20: CONSORT flow diagram cohort 3



5.2.3 School characteristics and summary of each school context

Nine of the eleven schools are located in the North East of England with two schools located in Derbyshire (Schools D and E). The randomisation of the pupils occurred at individual level, therefore no schools were involved as a control so no detailed analysis has been included. However, table 8 outlines the key characteristics of the schools.

Table 8: Characteristics of the schools

School	Level	Number pupils	KS1 / KS2	FSM (%)	EAL (%)	IDACI (%)
0	Secondary	799	28.8	18.5	2.9	18.3
1	Primary	123	2.9	5.5	7.7	15.8
2	Primary	234	0.0	9.8	10.7	1.3
3	Secondary	844	27.2	33.1	1.5	24.6
4	Primary	322	2.7	9.0	7.0	11.4
5	Secondary	798	26.7	48.7	1.1	64.9
6	Primary	141	2.2	63.0	3.0	85.9
7	Primary	112	2.5	38.9	0.6	43.2
8	Primary	118	2.6	13.1	5.1	4.5
9	Secondary	533	28.0	25.7	1.1	22.9
10	Primary	233	2.6	22.3	0.9	22.6

5.3 Outcomes and Analysis

As outlined in the statistical analysis plan (Appendix 3) the analyses of primary and secondary outcomes were undertaken on an 'intention to treat basis', meaning that all those allocated to treatment and control conditions in the randomisation are included in the final analysis, even if they were to drop out of the treatment (Torgerson & Torgerson, 2008). However, two schools in cohort 1 were unable to participate due to technical issues. The learners were not randomised into groups so the decision was made to exclude their data from the analysis, so it could be argued that the analysis was not strictly conducted under 'intention to treat'.

5.3.1 Descriptive analysis

Schools were asked during the planning stage to identify a time and day for the intervention to completed over a five-week delivery period: during a school lesson (not mathematics), at break time or after school. Nine schools (82%) selected during a school lesson and two schools (18%) selected after school as an enrichment activity for transition.

Appendix 9 includes effects sizes calculated from post-test scores (Cohen's d and Hedges' g) for 1:2 and 1:4 group sizes compared to 1:1 active control for cohorts 1, 2 and 3. The data was also modelled including the pre- and post-test using ANCOVA to calculate the effect sizes (Cohen's d and Hedges' g) for 1:2 and 1:4 group sizes compared to 1:1 active control. These are not included in the main analysis as the effect of school and pre-test data needed to be included into the model due to variations between schools and differences between the groups in pre-test scores.

Table 9 provides mean post-test scores for each school and the variance. The table demonstrates that the variance assumption of equal variance is not met with a variation greater than 2, this is 14.46 (12.15/ 0.86).

Table 9: School post-test data

School	Sample	Mean Post-test score	Variation (SD)
0	35	14.97	3.03
1	11	12.82	10.56
2	24	9.08	9.64
5	14	11.36	2.71
6	23	6.70	2.04
7	12	6.83	12.15
8	12	9.50	8.09
9	8	14.62	0.84
10	12	6.16	5.61

Schools 3 and 4 were excluded from the analysis due to IT issues preventing the intervention taking place in their respective schools. As randomisation had not occurred, the decision to remove the schools from the analysis was made. This could be argued that this is not intention to treat, however as the participants were not allocated to groups they have been removed from the primary analysis.

5.3.2 Cohort 1 primary analysis

Table 10 provides the pre-test and post-test mean scores for the active control (1:1) and group sizes 1:2 and 1:4. The mean pre-test scores for 1:2 (6.12) and 1:4 (5.79) differ from the active control 1:1 (7.68). The initial randomisation of tutees by blocking based on performance on the pre-test aimed to ensure an equal spread of ability per primary school, however variation between schools and combining peer tutors into the analysis (where simple randomisation had been used) has resulted in some imbalance.

Table 10: Pre- and Post-test data for group cohort 1

	1:1 (Active control)	1:2	1:4
Sample size	n= 25	n= 17	n= 14
Mean (variance) Pre-test scores	7.68 (28.98)	6.12 (23.38)	5.79 (31.10)
Mean (variance) Post-test scores	12.92 (10.91)	12.35 (16.12)	10.57 (21.80)

Intra-cluster coefficients and correlation (ICC)

The intra-cluster correlation coefficient (ICC) was estimated using the data from the assessments to be 0.15.

Multi-level model

The primary outcome analysis was conducted using a linear mixed effects model fit by REML (Restricted Maximum Likelihood). In the model, the school was random effects with pre- and post-tests as fixed. The school is random as it is assumed the effect varies by school. Pre- and Post-tests are fixed as it is assumed to be measured without measurement error and the variable used in the study contains all or most of the variables values in the population (Kasim, Xiao, Higgins et al., 2014).

The following assumptions in the model for multiple group comparisons are:

1. Independence of observations within groups
2. Independence of observations between groups
3. Normal distributions of observations
4. The groups share a common variance

The rationale for using a multi-level model (MLM) is the assumption independence within schools is most likely to be violated. When considering the normality assumption, normal t-based procedures are robust to it, which means, even if the normality assumption is not met and sample is sufficiently large enough, the procedure would produce more or less the same results. Yet, they are sensitive to the violation of equal variance assumption. For example, in smaller sample (n), it is more likely to reject the null hypothesis too often (false positive, type 1 error). In larger n, it is more likely to under-reject the null when the alternative is true (type II error).

The total variability is used because those derived from between and within variances can be biased, representing the most conservative approach and least likely to result in false positives compared to using within or between variability (Kasim, Xiao, Higgins et al., 2014).

The effect sizes and respective 95 percent confidence intervals were calculated, these are reported using Hedges' g rather than Cohen's d. The reasoning for selecting Hedge's g is that the method both pool variances on the assumption of equal population variances. As the method to calculate g pools variance using n-1 for each sample instead of n, it provides a better estimate than Cohen's d with smaller sample sizes (Grissom & Kim, 2005).

Table 11: Analysis of primary data Cohort 1

	1:2	1:4
Effect Size (Hedges' g)	0.01 (-0.99 to 1.01)	-0.14 (-1.34 to 1.06)
School Standard Deviation Intercept (residual)	2.56 (1.49)	2.57 (2.11)
Intra-class correlation	0.25	0.4
Pre-test		
Value	0.33	0.32
Standard error	0.12	0.13
DF	36	35
t-value	3.22	2.57
p-value	<0.00	0.01
Post-test		
Value	-0.51	-0.51
Standard error	0.43	0.43
DF	35	35
t-value	-1.2	-1.2
p-value	0.25	0.25
Total sample size	40	41

Table 11 reports the effect sizes of the groups 1:2 and 1:4 compared to the active control 1:1. The effect of group size 1:2 shows a slight difference of 0.01 (95% CI: -0.99 to 1.01) in the mathematics performance as defined by the data handling assessment, whereas the group size 1:4 is -0.14 (-1.34 to 1.06). The data also indicates a large intra-class correlation of 0.4 (1:4) when compared to 0.25 (1:2). Both the pre-tests for 1:2 and 1:4 were statistically significant at $p < 0.01$, justifying their inclusion in analysis as a covariate in the model.

5.3.3 Cohort 2 primary analysis

Table 12 provides the pre-test and post-test mean scores for the active control (1:1) and group sizes 1:2 and 1:4. The mean pre-test scores for 1:2 (3.94) and 1:4 (3.25) differ from the active control 1:1 (2.77). The initial randomisation of tutees by blocking based on performance on the pre-test aim to ensure an equal spread of ability per primary school, however variation between schools and combining peer tutors into the analysis (simple randomisation had been used) has resulted in a slight imbalance.

Table 12: Pre- and Post-test data for group cohort 2

	1:1 (Active control)	1:2	1:4
Sample size	n= 16	n= 12	n= 9
Mean (variance) Pre-test scores	3.94 (7.26)	3.25 (7.48)	2.77 (31.10)
Mean (variance) Post-test scores	8.88 (8.52)	8.83 (7.61)	7.22 (4.94)

Intra-cluster coefficients and correlation (ICC)

The intra-cluster correlation coefficient was estimated using the data from the assessments to be 0.15.

Multi-level model

The primary outcome analysis was conducted using a linear mixed effects model fit by REML (Restricted Maximum Likelihood). In the model, the school was random effects with pre- and post-tests as fixed.

Table 13: Analysis of primary data Cohort 2

	1:2	1:4
Effect Size (Hedges' g)	0.22 (-1.47 to 1.91)	-0.16 (-1.46 to 1.14)
School Standard Deviation Intercept (residual)	1.53 (2.10)	1.41 (0.84)
Intra-class correlation	0.65	0.27
Pre-test		
Value	0.40	0.75
Standard error	0.22	0.20
DF	24	21
t-value	1.84	3.85
p-value	0.08	<0.00
Post-test		
Value	0.69	-0.28
Standard error	0.59	0.30
DF	24	21
t-value	1.16	-0.93
p-value	0.25	0.36
Total sample size	28	25

Table 13 reports the effect sizes of the groups 1:2 and 1:4 compared to the active control 1:1. The effect of group size 1:2 shows a slight difference of 0.22 (95% CI: -1.47 to 1.91) in the mathematics performance as defined by the data handling assessment, whereas the group size 1:4 is -0.16 (-1.46 to 1.14). The data also indicates a large intra-class correlation of 0.65 (1:4) when compared to 0.27 (1:2). The pre-test for 1:2 was just above the threshold for statistical significance, however the pre-test for 1:4 was <0.00 and therefore justifying their inclusion in analysis as a covariate in the model.

5.3.4 Cohort 3 primary analysis

Table 14 provides the pre-test and post-test mean scores for the active control (1:1) and group sizes 1:2 and 1:4. The mean pre-test scores for 1:2 (8.06) and 1:4 (5.93) differ from the active control 1:1 (9.00). The initial randomisation of tutees by blocking based on performance on the pre-test aimed to ensure an equal spread of ability per school for learners, however variation between schools and combining peer tutors into the analysis (simple randomisation had been used) has resulted in some imbalance.

Table 14: Pre- and Post-test data for group cohort 3

	1:1 (Active control)	1:2	1:4
Sample size	n= 25	n= 18	n= 15
Mean (variance) Pre-test scores	9.00 (21.92)	8.06 (27.82)	5.93 (14.35)
Mean (variance) Post-test scores	11.04 (18.79)	10.17 (20.50)	8.67 (13.95)

Intra-cluster coefficients and correlation (ICC)

The intra-cluster correlation coefficient was estimated using the data from the assessments to be 0.15.

5.3.3 Multi-level model

The primary outcome analysis was conducted using a linear mixed effects model fit by REML (Restricted Maximum Likelihood). In the model, the school was random effects with pre- and post-tests as fixed.

Table 15: Analysis of primary data Cohort 3

	1:2	1:4
Effect Size (Hedges' g)	0.05 (-0.66 to 0.56)	0.17 (-0.47 to 0.81)
School Standard Deviation Intercept (residual)	1.18 (5.50)	1.12 (3.64)
Intra-class correlation	0.0	0.0
Pre-test		
Value	0.86	0.90
Standard error	0.04	0.04
DF	36	33
t-value	23.06	20.72
p-value	<0.00	<0.00
Post-test		
Value	-0.06	0.20
Standard error	0.36	0.20
DF	36	33
t-value	-0.16	0.98
p-value	0.87	0.33
Total sample size	43	40

Table 15 reports the effect sizes of the groups 1:2 and 1:4 compared to the active control 1:1. The effect of group size 1:2 shows a slight difference of 0.05 (95% CI: -0.66 to 0.56) in the mathematics performance as defined by the data handling assessment, whereas the group size 1:4 is 0.17 (-0.47 to 0.81). The data also indicates minimal intra-class correlation of 0 for both the groups. Both the pre-tests for 1:2 and 1:4 were statistically significant at $p < 0.01$, justifying their inclusion in analysis as a covariate in the model.

5.4 Cumulative data-analysis

The primary analysis of the three cohorts involved in the aggregated trial (n=151) is not sufficient to make inferences regarding the effectiveness of the group sizes to answer the research questions relating to the effective of 1:4 and 1:2 online group sizes compared to a 1:1 control. Table 16 shows that in all three cohorts, 1:2 provided a positive effect sizes of 0.01, 0.22 and 0.05. The data demonstrates that 1:2 peer tuition has a similar impact to 1:1 peer interactions within the context of the sample analysed.

Table 16: Cumulated trial data

Cohort		1:2	1:4
1	Effect Size (Hedges' g)	0.01 (-0.99 to 1.01)	-0.14 (-1.34 to 1.06)
2	Effect Size (Hedges' g)	0.22 (-1.47 to 1.91)	-0.16 (-1.46 to 1.14)
3	Effect Size (Hedges' g)	0.05 (-0.66 to 0.56)	0.17 (-0.47 to 81)

When the data is aggregated in the form of a prospective cumulative meta-analysis, the plots show demonstrate the following.

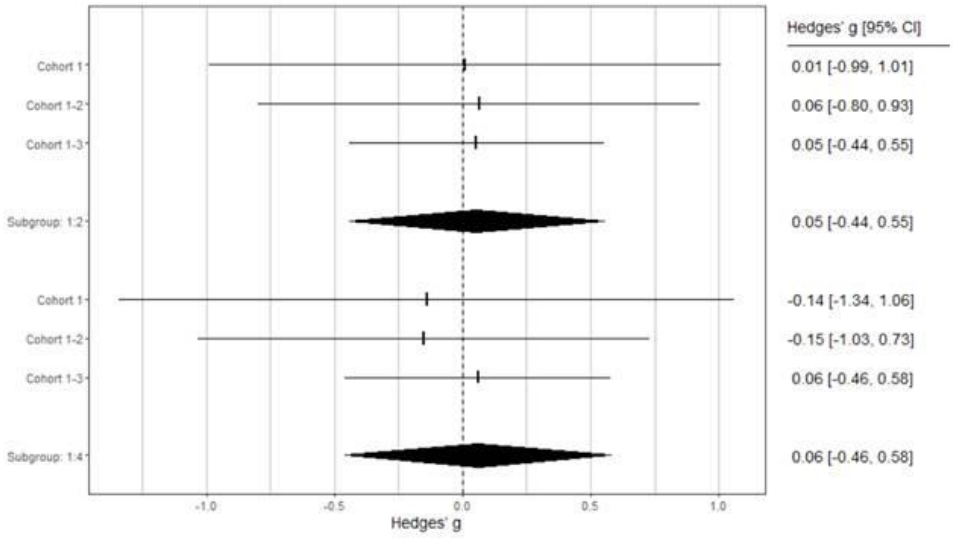


Figure 21: Cumulative meta-analysis of online peer tutoring aggregated data.

The subgroup 1:2 has an effect size of 0.05 (-0.44, 0.55) and 1:4 has an effect size of 0.06 (-0.46, 0.58) when compared to the active control 1:1. In each of the cohorts the effect sizes are fairly stable, particularly for the 1:2 group size. In order for any generalisations to be made for the research questions for the relative effectiveness of one to two and one to many online peer tuition group sizes, additional cohorts are required with schools selected randomly from across the UK and sufficiently powered to draw such conclusions.

5.5 Exploratory analysis

The statistical analysis plan in Appendix 3 explains that intention to treat (ITT) was intended to be used throughout the analysis of the data. In the first cohort, two schools were not able to participate due to IT issues. As the learners and tutors were not randomised to groups, the decision was made to remove these from the analysis. Therefore, the analysis should not be considered to be completed under ITT conditions.

A pragmatic decision was made in agreement with the school for the learners and tutors to still complete the pre- and post-assessments to provide exploratory data as a small comparison group. The following analysis is not outlined in the SAP and the finding will not be used in the discussions section of this thesis as these are outside of the intended research. Therefore, no generalisations can be made regarding the effectiveness of online peer tutoring. The results of the exploratory analysis are in Appendix 4.

5.6 Process evaluation (Part 1)

The process evaluation had two main purposes. Firstly, it assessed the fidelity of the delivery of the online peer tutoring, as this is a fundamental prerequisite for the intervention. Secondly, it addressed the feasibility of conducting an aggregated trial for investigating the effectiveness of online cross-age peer tuition group sizes across the transition boundary between primary and secondary schools.

The next section outlines the key components of the process evaluation, providing context to the implementation of the intervention prior to addressing the process evaluation questions outlined in Chapter 4.

5.6.1 Implementation

At the start of each cohort, planning meetings were arranged with partner schools. The IT requirements were discussed and IT testing completed with each school. The IT testing involved opening the relevant ports for the firewall settings, testing the sound settings on the computers and confirming schools had access to headsets with microphones. Ten schools (90%) did not have access to headsets with microphones, these were provided by the researcher in advance of the intervention starting in each cohort.

Unfortunately, two schools (D and E) from the Derbyshire region were unable to clear the firewall settings prior to the start of the cohort 1 delivery. The researcher visited the school on three occasions and liaised with the IT team at TUTE and the independent IT provider of the primary school. As the service had been outsourced to an American company at the start of the academic year, we were not successful in opening the relevant ports for the interactive classroom to function correctly. As the pre-test had already been completed, the school agreed to complete the post-assessment and act as a retrospective control group. The data is not included in analysis has this was outside of the

statistical analysis plan. Therefore this means that the study was not strictly ITT and this will be discussed further in the limitations section of this thesis.

The delivery of the intervention mainly occurred in either specialised computer rooms in secondary schools or a standard classroom using laptops in primary schools. Typically, pupils arrived 10 minutes before the start of the online peer tuition lessons, to start up the computers and log into the TUTE online platform. A difficulty encountered in each of the three cohorts involved the availability of computer rooms in secondary schools, with room changes sometimes occurring at the last minute due to double bookings of the classrooms.

5.6.2 Length and timings

The online peer tuition lessons were 30 minutes in duration. As part of the pre-planning meeting with schools, peer tutors and learners were advised to arrive at the computer room 10 minutes before the lessons commenced to log into their school computers. A logistical issue encountered in previous research (Harrison, 2015) involved mentors arriving just before the start time, arriving late into the online classroom due to the length of time taken to load up their school pupil settings and to access the internet. The addition of the direct instruction to arrive 10 minutes prior to the start reduced the occurrences of delayed starts to the mentoring. Lesson observations noted that the majority of peer mentors were able to log into the TUTE platform before the learners arrived, maximising the 30 minutes available for the peer tuition. In the first few weeks in school A, two peer tutors had to be reminded to attend as they had forgotten about the online peer tuition. In order to mitigate this issue in future cohorts, a reminder system was implemented to inform form teachers to remind peer tutors on the morning of the lesson.

A focus group question asked peer mentors about the length of the lessons and if they felt they had sufficient time to deliver the content. Mentors agreed that the length of time was sufficient to explain the concepts and allow their learners to answer the questions.

5.6.3 Technical set-up

Initial set up

The successful implementation of online peer tutoring is dependent on the school IT infrastructure and personal within these institutions. Secondary schools have a dedicated IT team and infrastructure manager, therefore changes to firewall settings are relatively straight forward. Unfortunately, primary schools are smaller institutions and often outsource IT support as it is not financially viable to employee dedicated IT support. The aggregated trial had a higher proportion (63%) of primary schools compared to secondary schools (37%), resulting in logistical time delays as IT support was either weekly or fortnightly. Accordingly, a 6-week implementation window (one school half-term) was allocated between cohorts to test the IT systems.

The TUTE online platform requires IT testing using TUTE technical support and an IT technician in each school. Three schools encountered problems with the testing of TUTE, one school had two separate firewalls requiring IT support each lesson to disable these and the other school had issues with audio due to the software installed on computers. These issues were managed throughout the project and did not impact on the delivery of the lessons.

On-going technical issues

Ongoing technical issues were isolated to secondary schools, often as a result of damage to computers in dedicated computer rooms. Often, chewing gum had been placed in the headset sockets or keys had been changed location on the keyboard. In order to minimise the risk of the headset sockets being unusable, peer tutors were directed to connect the headsets to the sockets at the back of the computers as these were less accessible and unlikely to be damaged.

An issue in one particular primary school (2) involved a weak wi-fi signal. The school used laptops connecting to wireless wifi and on some occasions the lessons would sometimes freeze and lose connection. This only occurred twice but it did disrupt the learning of the pupils involved, as it

required them to log out and load up the lesson again. After speaking with TUTEIT technicians, it was advised to try to use computers connected directly to the internet.

5.6.4 Online delivery

The lesson observations, focus groups and teacher interviews demonstrate that the learners and mentors had a positive experience of the online peer tuition programme. A reoccurring theme across the lesson observations in all three secondary schools participating in the research was the positive attitude to engaging in peer teaching using the online classroom.

The focus groups with Year 8 peer tutors also found generally positive views.

> *"I found the online tutoring really helpful as it made me think about how I solve the questions and being able to explain this has help me."*
>
> *"I usually talk with my friends online using my PlayStation, I've not done this for maths before but it was fun."*
>
> *"I liked the ability to mute the primary students when I was explaining something and then they could click on a hand to ask a question."*
>
> *"I really enjoyed teaching pupils from my old primary school."*

Year 6 tutees also responded positively to the online delivery of the mathematics content.

> *"I really find pie charts hard but X showed me how she does this and it makes sense."*
>
> *"It was nice to get to know an older pupil in the big school, they were really friendly and helped me with my maths."*
>
> *"I don't like maths when we do this in books but this was fun."*

One of the tutees commented that she enjoyed the experience of learning from another child online rather than face to face, and when asked why she said:

> *"I sometimes don't like to ask questions in the whole class as I don't want my friends to know that I cannot do something. My peer tutor is different so I don't mid telling them I don't know how to do it."*

When asked why this might be, she said:

> *"I don't know….. I think it is because they are not in my class and are not a teacher as well. It is easier to tell them I cannot do a problem."*

All the pupils in the focus groups have communicated online with someone they know through gaming, chat or social media. The most common communication was through game stations or mobile phones for the peer tutors, whilst tutees were less likely to own a mobile phone device so this was their family tablet or game station.

5.6.5 Behaviour and attitudes

The lesson observations of the peer mentors had almost no off-task behaviour when the peer tutors were mentoring in all three cohorts. The only instance of this occurred when the learners were late to log into the lessons. Teacher A commented:

> "*As soon as the headsets go on they become focused on delivering the maths lesson and you can hear almost no conversations that are not related to the task they are doing…. I might take the headsets into the classroom to see if this helps them focus!*"

Throughout the three cohorts the lesson observations consistently highlighted the lack of off task conversations. Furthermore, the quality of the conversations was also consistently high with mentors explaining how they work out problems using the correct mathematical terminology.

A limitation of the lesson observations relates to the lack of resources in the aggregated trials, as a single researcher could only be present at one venue at a time. As the IT issues were most likely to be at the secondary schools and IT support staff could not attend these lessons, the decision was made to focus observations in this particular setting. Therefore, it was not possible to observe the behaviour of the tutees in the classroom setting.

Interviews with the primary teachers highlighted the positive behaviour during the delivery of the online lessons. Teacher C explained:

> *"When the pupils put the headsets on it was great to hear that they were talking maths with their tutor and actually asking the tutors to explain if they were stuck".*

Teacher G explained:

> *"X is usually trying to get up out of his seat every two minutes when we do maths in the classroom. He was able to sit for 30 minutes in one place which is a first!"*

A recurring theme across both the secondary and primary teachers is the focus on the maths content when the tutors and tutees used a headset.

5.6.6 Human investment

The initial set up requires time from the IT technicians in schools to change the firewall settings and complete an IT test with the TUTE IT team. In order for online lessons to be delivered in schools, the pupils require supervision either from a teaching assistant or teacher. In three of the schools (27%), teachers supervised the lesson and teaching assistants were deployed in the remaining schools (73%). Teachers supervised mainly in secondary schools (100%) whereas primary schools predominantly used

teaching assistants. Ideally, the online peer tutoring does not require teacher supervision. However, these schools have limited teaching assistant resources so teaching staff were used to manage the intervention.

5.6.7 Support and commitment

TUTE supported the technical set up, creation of pupil accounts and arranging the lesson bookings for schools. When technical issues did occur such as pupils becoming locked out due to typing an incorrect password 3 times, the account was unlocked within 5 minutes of the incident occurring.

Within schools, teachers and teaching assistants were able to manage the programme throughout the three cohorts of the aggregated trial. Even for schools D and E, the school agreed to complete the post assessment when, as explained previously, the intervention could not be delivered.

IT support within schools was good during the testing of the firewalls, however due to staffing limitations the researcher had to provide technical support in the secondary schools if issues occurred with the headsets or computers. Due to the close involvement of the researcher in the efficacy trial, school support staff would need to be trained in troubleshooting these issues in future effectiveness trials.

5.6.8 Fidelity

The intervention was implemented as intended, and fidelity to the trial conditions was high. The researcher observed 196 lessons delivered by peer tutors (7 per school visit), with a conservation with each teaching assistant from the participating primary schools after each session (email or phone call) during the first two weeks to ensure they had sufficient support. All schools complied with the eligibility criteria and adhered to the 30-minute time allocation for the peer tuition lesson.

The measure used for fidelity consisted of three component scores relating to (a) how well the technology worked (b) the tutor attitudes to learning (c) content delivered in the 30 minutes. In each school visit each component was rated between 1 and 3, with 1 corresponding to 'poor', 2 corresponding to 'partially worked' and 3 corresponding to 'worked as planned'. Rather than an individual score for each online lesson, the score became an estimate for fidelity as a school and the average score calculated after the last lesson had been delivered. Therefore, each online tuition session had a fidelity score which ranged from 3 to 9. Lower scores correspond to lower fidelity and individual lesson observational comments were recorded as part of this process.

Table 17: Secondary school fidelity scores

School	Component A Technology	Component B Attitude	Component C Lesson content	Fidelity score
0	2.5	3	2.3	7.8
5	2.7	3	2.5	8.2
9	2.8	3	2.3	8.1

Discussions with all the schools indicated that the mentors or learners were not being taught data handling at the time of the trials and all the pupils in the trial were not receiving additional mathematics interventions involving this topic.

5.7 Process evaluation (Part 2)

My previous masters research (Harrison, 2015) demonstrated the basic feasibility for online cross-age peer tuition across the transition boundary. The purpose of this exploratory study's process evaluation was to overcome the logistical barriers encountered in the previous research and develop a scalable methodology to test a prospective cumulative meta-analysis as an approach in education. It is important to acknowledge that due to the nature of this exploratory study, the focus on the internal validity reduces the external validity of these findings due to the close involvement of the researcher in the implementation of the intervention (Bowen & Neill, 2013). Consequently, the findings are not generalisable outside the context of the efficacy study.

The process evaluation comprised of three stages. The first stage involved the evaluator working with school IT technicians and the TUTE technical support team to test the online collaborative platform in each of the participating schools. The second stage involved observing the online lessons delivered by Year 8 pupils over a period of 5 weeks in each of the participating schools. A fidelity measure was recorded on a scale of 0 to 9 for each lesson observed. The third stage of the process evaluation involved conducting focus groups with all the pupils participating as tutees or peer tutors, semi-structured interviews with teachers and IT technicians in schools.

Semi-structured interviews were conducted with teachers in five schools and were recorded using an audio device. The focus groups and interviews were transcribed then imported into NVIVO software. The data was analysed using an inductive approach to a Thematic Analysis based on the framework by Braun & Clarke's (2006) 'guide' to the six phases of conducting this process. A limitation with the analysis was that only one person coded the data, therefore interrater reliability was not possible. However, a coding framework was developed to ensure the method was applied consistently.

The process evaluation research questions which are set in in chapter 4 will now be revisited and answered using supporting evidence from the lesson observations, pupil focus groups and teacher interviews.

5.7.1 How feasible is it to implement an online cross-age peer tuition project across the transition boundary of primary to secondary school as an aggregated trial?

This exploratory study has replicated aspects of my previous study (Harrison, 2015) to demonstrate that the concept of online cross-age peer tuition is feasible between the transition boundary between primary and secondary schools. A limitation of the previous research involved the study running the trials in two schools at the same concurrent time interval due to the time restraints of the MA research. The current study involved the aggregation of three cohorts of the same trial to allow the data to form the evidence for a cumulative meta-analysis over a school academic year. The timeline of the three cohorts of schools are outlined in Table 18.

Table 18: Overview of the aggregated trial cohorts over the academic year 2015 - 2016

	Secondary school	Primary school	Academic term
Cohort 1	0 3	1 & 2 4	Autumn 2 Nov – Dec 2015
Cohort 2	5	6	Spring 2 Feb – March 2016
Cohort 3	0 9	7 & 8 10	Summer 2 June – July 2016

The study has demonstrated the feasibility of using small scale aggregated trials across a sequential time period using a specific protocol to enable replication of the individual studies.

5.7.2 How feasible and acceptable do teachers feel it is for using online peer tuition as a transition intervention?

Semi-structured interviews were conducted with three secondary teachers and three primary teachers over the duration of the data collection period. All teachers were positive towards the use of online cross-age peer tuition and the use of technology to engage the learners and tutors in mathematics.

Teacher F commented:

> *"We currently work with up to 19 feeder primary schools and it is often not possible to provide opportunities to schools that are further afield. The online delivery allows our students to work with primary students without having to organise transport, teacher cover and avoids the costs involved in the transport of pupils."*

Teacher A reinforced the flexibility of the online delivery:

> *"If we planned a traditional face to face peer tutoring activity, this would require the primary pupils to be transported to the school and this takes time and financial costs. As a one off this could be justified but as a weekly activity we would not be able to coordinate or afford this type of intervention. The online delivery means that the intervention is delivered in 30 minutes and the pupils (tutors and tutees) can return to their lessons."*

When asked if the online delivery enabled a more sustainable intervention, the teacher responded:

> *"Definitely, we have been able to implement an intervention over a half-term with minimal costs, paperwork and disruption to pupils. The attendance has been very high for both the learners and tutors and it has been great to see how the tutors have been able to explain concepts."*

When asked to expand on how this reduced paperwork, the teacher responded:

> *"If we have to take pupils offsite for an intervention or trip, we have to complete a health and safety risk assessment. As the intervention is online and the pupils do not leave the premises, I don't have to complete these".*

Another key theme emerging from the interviews was the behaviour of the pupils during the intervention. Teacher B explained:

> *"It has been great to listen to my pupils talk maths with the peer tutors, as a number of the learners chosen for the intervention struggle to concentrate in lessons. Throughout the programme of lessons all the pupils have been engaged and I hardly heard any off task conversations".*

When asked why she thought the pupils were more focused, the teacher responded:

> *"I think it is because they are wearing headsets and this blocks out other distractions, so they focus on the work that they are doing. I often find that the pupils can get distracted easily in maths lessons so it has been great to see them engage with their peer tutors".*

This theme regarding behaviour in the secondary school teachers is also clear, with all three teachers highlighting the positive attitude and behaviour of the tutors who had been selected. Teacher J explained:

> *"When I selected the pupils to become peer tutors, I specifically chose X and Y as they often lack confidence but they have the ability. They have really engaged in being a peer tutor and it has been great to hear them explain the concepts to the primary pupils. I have been really impressed with their communication skills as I often do not see this in the classroom environment".*

Teacher F reinforced the positive attitude of the tutors towards the programme:

> *"The tutors actively chased me for the lesson content each week if I forgot to hand this out at the end of the lesson they had delivered. If this had been homework this would not be the case."*

When asked to explain why this might be, the teacher responded:

> *"I think it is because they want to make sure that they know what they are teaching to their primary students, so they don't look like they do not know what they are doing in front of their younger peers".*

The teacher expanded on this further:

> *"I think this is great because you know that they are reinforcing their own learning by reading through the slides and I have had a few pupils ask me to go through a concept before they delivered their own lesson."*

When asked about how feasible it is for schools to use online small cross-age peer tuition as an intervention, all six teachers were positive about the approach of the intervention. The flexible

delivery of the online delivery of the intervention was a key theme linked to the feasibility of running this type of intervention in the future.

Teacher F explained:

> *"If we could access a platform that allowed us to allocate peer tutors to learners in our primary schools, I see this as a very powerful intervention. Due to the way the intervention works, we could run multiple programmes with a number of feeder primary schools with minimal planning compared to a traditional face to face programme."*

The teacher was asked to expand on why they see this as a powerful intervention:

> *"Basically you are getting a two for one intervention, as the programme reinforces the knowledge of the pupils who are selected as a peer tutor and also the learners in the primary schools. I also believe that the tutors are also developing their softer skills such as communication and this is invaluable".*

Cross-age peer tutoring strategies demonstrate a positive impact on attainment for the tutees as well as the tutors (Jun, Ramirez, & Cumming, 2010; Leung, 2015). The empirical data from the current study supports the teachers views, as the mean pre-test tutor score (10.93) and post-test tutor scores (14.04) demonstrate a potential positive impact on tutor attainment.

However, one negative theme all across the teachers' responses was a concern about the potential cost of using a platform to deliver the interventions.

Teacher A explained:

"It has been great that platform that we have used has been provided free for the intervention for the research, however a licence fee might not be affordable if this was excessively expensive".

The theme of costs were also highlighted by the primary school teachers, with teacher B explaining:

"It would not be possible for our school to take part if we had to fund the intervention, as our budgets really stretched already. I would hope that the secondary schools could use this as part of their transition budgets".

5.7.3 What are the barriers to the successful implementation of online cross-age peer tuition as a transition intervention? How can these be minimised in future aggregated trials?

The potential barriers to the successful implementation of online cross-age peer tuition programme can be classified into three main points:

1) Technical – A fundamental component to the success of the intervention is the quality of the IT infrastructure within the schools. IT infrastructure includes the quality of internet (bandwidth), firewall permissions, compatible computers and maintenance of devices in a school environment.
2) Logistical – The time required to deliver the lesson in school, access to computer rooms, headset availability and staffing requirements.
3) Financial – Sufficient budgets within schools to purchase the intervention.

In order to minimise the three main barriers to implementation for future aggregated trials involved in cross-age peer tutoring, each barrier can be reduced through the following:

Technical

Unfortunately, due to the online delivery of the intervention, technical barriers can be frustrating if these are not managed or tested prior to the start of the intervention. In order to reduce the risk of

technical problems preventing the delivery of the online peer tutoring, the following implementation steps were used:

1) Firewalls – All schools have a firewall as part of their cyber security processes and these are required to be tested prior to the intervention starting. In each school, the IT technician was contacted 6 weeks prior to the intervention starting with the details for the correct ports to be configured and the website whitelisting to allow users to access the online platform. As primary schools do not have full time IT technicians, the timescales for the process to be completed ranged from one to three weeks. Unfortunately, in one instance with schools D and E it was not possible for the primary school to make the relevant changes in time (even with a 6-week timeline).
2) Testing – Prior to the intervention starting, it is advised that all the computers that will be used for the intervention are tested to ensure the sound is working. As best practice, if these computers are marked on a plan so that pupils use the same computer each week. The majority of the technical issues related to pupils using sound ports at the front of computers, rather than the back. Unfortunately, the ports on the front are often damaged with objects such as chewing gum or pencils, so using the rear ports minimises this risk.
3) Headset permissions – All schools have different user permissions depending on their policies. If headsets are plugged into the PC and these are not set as default, then they are unable to change the settings to allow the microphones to work. It is important to check the settings as part of the testing process to ensure the sound works.

Logistical

The main logistical issue for schools involved locating a computer room during the school day which was not used for lessons. In school environments, it will not be possible to minimise this particular barrier to the often-limited availability of computer classrooms. The second main barrier for logistics related to access to headsets with microphones. In order to facilitate the online peer tutoring,

headsets with microphones are required. A potential barrier for future trials is that schools often have headsets but these do not have microphones, so for the current study these were purchased by the researcher. It is important to ensure that school have the correct headsets as when questioned all school replied that they had these in the initial meetings, when actually they had headsets without microphones.

Financial

The platform and lessons were provided at no cost to schools participating in the initial feasibility study, however future usage would require a licence fee with the platform provider. In order for the intervention to be scalable, a cost-effective online software would need to be specifically built for school usage or for schools to test free online platforms such as Google hangouts or Skype. However, these commercial platforms are not designed for online collaborative learning.

5.7.4 What are the views of the intervention of the peer tutors, tutees and teachers?

Tutee views

The focus groups with Year 6 pupils found positive views.

> *"I really enjoyed working with my Year 8 tutor…. they were able to show me how to work out pie charts as I did not know how to do this before."*

> *"My tutor was really good and they have helped me learn loads. I feel more confident with my maths now."*

> *"It felt just like talking to a friend my Xbox but it was for maths, it was really good and I learned a lot from my tutor."*

"I loved learning from another pupil rather than a teacher, they can explain it better."

A theme across all the Year 6 focus groups was the ability to speak with pupils who were at secondary school and ask questions about the differences between there schools. When asked why they enjoyed this, one pupil said:

"I know when I go to secondary school it will be really different from our school. I enjoyed asking X about how they found starting Year 7 as I know they went to our school before. Having different teachers for different subjects will be really good."

Another pupil followed this by saying:

"I asked my tutor about the school lunches and you have to charge your finger with money using a machine and you have lots of different choices of food."

When asked as a group which questions they had asked their peer tutors about in the first get to know you lesson, the three main topics were homework, is it easy to make friends and what are the teachers like.

Research shows that he transition from primary to secondary school in the UK, and its equivalent elsewhere, has been depicted as 'one of the most difficult phases in pupils educational careers' (West et al., 2010; Zeedyk et al., 2003). It is almost universally agreed that the majority of children express some form of concern or anxiety prior to the transition from primary to secondary schools (Chedzoy & Burden, 2005; Graham & Hill, 2003; Measor & Woods, 1984; Shepherd & Roker, 2005). The opportunity for the tutees to interact with their peers prior to transition should reduce the anxiety, however no generalisations can be made as this was not the purpose of the research.

A similar theme to emerge linked to the transition the Year 6 pupils involved confidence, either in the subject of maths or the ability to speak with another older student at the secondary school.

> *"I could not work out pie charts before my peer tutor showed me how to do these. What I really liked was that they could not do this at the start of Year 6 but by the time they did their SATs they said it was easy. I feel happier that I can work these out now".*

> *"I feel like I have a friend at the school already as X was really helpful in the lessons. It would be really nice if we could meet our peer tutors in person before we go to the big school".*

> *"My tutor has helped me learn maths in a fun way, it was like being on my PlayStation but in school. I was able to answer questions on the second test that I could not do before."*

> *"It would be really good if we could have another lesson where we could talk about being in secondary school. I enjoyed the maths but being able to ask questions about what is like has made me feel less nervous about going".*

Confidence was an associated factor identified by the tutees, with students agreeing that they were more confident in asking questions in an online environment compared to traditional peer tutoring environments.

> *"I feel more confident talking online rather than in a class, as I don't want my teacher and friends to know that I cannot do something".*

This supports research that online learning alters the dynamics of the interactions between the tutor and tutee, with speculation that it allows the tutees to become more involved in their learning and to initiate more questions than face-to-face (Jones et al., 2006).

The focus groups demonstrated that the intervention had provided a positive learning experience for engagement in maths and bridging the transition between primary and secondary schools. The focus of the feasibility study involved the methodological development of using aggregated small scale RCTs to link to a cumulative prospective meta-analysis. In future studies, using questionnaires as a measure of engagement and confidence for the transition will improve pupils on the design of the process evaluation.

Tutor views

A theme emerging from the peer tutors was communication and familiarity. The majority of students are comfortable communicating with fellow peers online, as this is part of their daily routines through the use of mobile phones, tablets (facetime) and gaming platforms. All the peer tutors involved in the intervention regularly use online communication with either friends or family.

> *"I speak with my friends every night when we play FIFA and other games on my Xbox so talking about maths was not really any different".*

The use of gaming platforms was a consistent theme in all the Year 8 focus groups, with the majority using this as a way of communicating with friends outside of school. As the Internet has become more accessible and affordable, this is changing the way individuals can communicate through virtual environments. Content analyses of social interactions within online gaming spaces have revealed that

emotional communication vastly predominates task-orientated conversations (Kowert & Oldmeadow, 2013; Peña & Hancock, 2006). The accessibility of gaming platforms and virtual spaces for collaboration in gaming environments has found that up to 75% of game players reported having "good friends" within their gaming communities (Cole & Griffiths, 2007), highlighting the social bonds formed through communicating online.

Even though the programme lasted only half a term, learners had mentioned that they had formed friendships with tutors during the lessons. As an intervention the ability for tutors to form relationships with tutees prior to the transition should reduce the anxiety and stress often linked to the transition between primary and secondary schools.

Teacher views

An emergent theme from the interviews with teachers was the ability for the online delivery to facilitate peer tutoring by removing the logistical barriers that would be encountered when using a similar face to face approach. Teachers felt that traditional cross-age peer tuition would not be feasible to deliver across the transition boundary between primary and secondary schools. The logistical barriers identified reinforced the views from previous research (Harrison, 2015) and included time (travel and delivery), paperwork (Evolve Health and Safety), cost (staffing, transport, administration) and practicality of delivering this over a half term period.

All the secondary schools involved in the feasibility study had more than 8 feeder primary schools, with a number of these schools spread across a wide geographical area. In traditional cross-age peer tutoring using face to face contact, all the teachers agreed that the logistical and financial barriers would prevent them from running a programme over a half term period.

Topping et al., (2003) conducted a cross-age peer tutoring programme for mathematics between primary and secondary pupils. The short nature of the intervention was addressed in the limitations with recommendations to trial a longer programme. However, they acknowledge this might not be feasible or sustainable due to overcrowded curriculums and logistical barriers. It is important to acknowledge that the study was conducted in 2003, so computing provision within schools is likely to be more flexible and available.

An emerging theme around the logistical barriers were a potential concern if the intervention was to be deployed as a long-term solution. The issue was highlighted by teacher A:

> *"The most difficult aspect of the programme was finding a computer room that was available to use for the intervention. Our school is moving towards tablets rather than dedicated computer rooms, so these are like gold dust to find in our school".*

When prompted if they could suggest a solution to the rooming constraints, the teacher responded:

> *"The online platform has to be used on PC and as these are limited in our school, it would make more sense if the pupils could use mobile devices such as tablets. This would also make it easier for pupils to draw and write on, as using the mouse to draw is difficult for the learners."*

In the primary schools, all three teachers interviewed highlighted that the intervention required a teaching assistant to supervise the intervention. An unintended consequence of the interventions is a possible spill over effect on eth remaining pupils in the class. As twelve pupils were withdrawn for the intervention, the class size had been reduced student to teacher ration which could provide a positive spill over effect.

Teachers were also positive in the engagement of learners with the online peer tuition, reinforcing the views of the learners from the Year 6 and Year 8 focus groups.

5.7.5 What are the staff perceptions of the current and possibly sustained impact of the intervention on children's educational attainment?

The teacher's perceptions of the current and possibly sustained impact of the intervention were positive. An emergent theme related to the perceived impact on tutors attainment involved the tutors reading through the lesson content in advance and asking teachers to clarify concepts if they were unsure.

> *"The tutors actively chased me for the lesson content each week if I forgot to hand this out at the end of the lesson they had delivered. If this had been homework this would not be the case."*

> *"I think the intervention has had a positive impact academic impact on the students I selected as tutors, as a number of times I had pupils ask me in a lesson prior to the intervention to go through a concept as they were not sure which was the best way to explain. Then seeing these students explain this in the following weeks online lessons showed that they had fully grasped the concept".*

5.7.6 What are their suggestions for change if the intervention was to be more widely implemented?

The emerging themes from the interviews with teachers were around the potential logistical barrier of access to computer rooms on a regular basis, the ability for the programme to offer more subjects and if a programme could be designed based purely on transition (pastoral) rather than academic for pupils with educational special needs. Firstly, all six teachers identified a move towards their schools purchasing iPad or tablet style devices, even in primary schools. In secondary schools, traditional IT rooms had been converted to classrooms due to the purchase of class sets of mobile tablet devices.

"The most difficult aspect of the programme was finding a computer room that was available to use for the intervention. Our school is moving towards tablets rather than dedicated computer rooms, so these are like gold dust to find in our school".

"If the software worked on mobile tablets we would be able to deliver this anywhere in the school and rooming would not be an issue."

Habler, Major & Hennessy (2016) hint at the possibility that introducing tablets is reducing the use of desktop computers in computer labs, reinforcing the point made by the teachers. Yet the evidence base for the usage and impact of mobile learning is fragmented with very few rigorous studies.

The second them emerging from the interviews focused on the ability of the online peer tutoring programmes to offer support for additional subject areas such as Modern Foreign Languages, English and Science.

"*I am sure that if our MFL department see how the lesson works they will want to use this for transition as this will really benefit the speaking skills of the tutors".*

"I could see online peer tutoring working well in subjects such as French and English. The ability to talk through your work with a peer would be a great learning experience for the learners and tutors".

The third theme to emerge from the interviews with the teachers revolved around the first lesson structure, allowing tutees to ask their tutors about their own transition to secondary school.

"After the first lesson my pupils were buzzing about going to secondary school and when they returned to their class they were talking to their peers. It was a shame that the students who would probably benefit the most for this non-academic lesson could not participate."

When asked if they could expand on this, the teacher commented:

"The maths programme was designed to stretch the more able pupils, yet the transition lesson would be really useful for our SEN children. If the online tutoring could be used to buddy up pupils in the term before the transition then I think this could make a positive impact on these pupils."

Mowat (2019) echoes the view of the teacher for focusing the transition intervention on vulnerable students by explaining that "for children who may already be struggling with the day-to-day realities of school life, these anxieties, fears and stresses may be compounded". Therefore it is important for future research to explore how online peer tutoring can support SEN pupils in the transition between primary and secondary schools.

5.8 Conclusions

The headline finding is that it is certainly feasible to implement a small scale-aggregated trial in a sequential timescale for an academic online peer tuition intervention across the transition boundary between KS2 and KS3. It is important to acknowledge the close involvement of the researcher in the organisation and successful delivery of the programme. It is plausible that without this support the teachers involved may have struggled to implement the intervention.

5.8.1 Limitations

The design of the study involved three cohorts of schools completing data collection over one academic year. Due to logistical limitations and the high-level support required for a feasibility efficacy study, the sample size is not sufficient to make generalisations from the analysis of the data for the group size effectiveness. Consequently, the study is limited in the generalisations that can be inferred within the context of this study.

The development of the assessment for the trial is a limitation due to not using an independent assessment measure that had been used in previous research (Harrison, 2015). However, with the Rasch analysis for the assessment reporting an overall person reliability of 0.86 and an item reliability of 0.96, a pragmatic decision was made to use this as an alternative to the standardised InCAS assessments previously used. Furthermore, Cheung & Slavin (2016) found that effect sizes for experimenter-made measures were much larger than studies using standardised instruments. Even after adjustment in their study, experimenter-designed instruments were twice the size of effect sizes compared to standardised tests. If financial constraints were not a limitation, then independent standardised assessments such as INCAS or other standardised tests of mathematics would be purchased.

Technical constraints one primary school resulted in two schools being unable to participate due to the nature of the online delivery of the intervention. A limitation with online interventions is the requirement for the IT infrastructure to work. The current study allowed 6 weeks for the clearing of firewalls and testing in schools in each of the three cohorts, yet it was still not possible to prevent the firewall issue from impacting on the study. (TUTE developers were not involved in this evaluation of the intervention which means that there was no conflict of interest between the technology company and the intervention design.)

5.8.2 Key feasibility conclusions

The study demonstrated the ability for a single researcher to conduct a series of small scale randomised controlled trials sequentially over one academic year to allow the aggregation of the data for a prospective cumulative meta-analysis.

5.8.3 Impact evaluation conclusions

The main objective of this study was to explore whether testing the relative effectiveness of online peer tuition across varying group sizes 1:2 and 1:4 in mathematics ability against those of pupils in the active control group (1:1). The requirements for the trial were maintained throughout, though with two schools dropping out of the study due to IT barriers, the online peer tuition intervention was delivered over a 6-week period with all pupils participating in the intervention in the other schools. Consequently, the Hawthorne effect was not potential source of bias in the results.

Even though the trial included three cohorts, the sample size is small, therefore the effect sizes must be treated with caution. The overall finding suggest that the 1:2 group size had a positive effect size in all three cohorts; 0.01 (-0.99 to 1.01), 0.22 (-1.47 to 1.91) and 0.05 (-.66 – 0.56) when compared to the active control (1:1).

5.8.4 Process evaluation conclusions

The process evaluation confirmed that the intervention was able to be delivered with a high degree of fidelity. Consequently, if the trial was conducted as an effectiveness study without the high level of researcher involvement, potential IT and logistical barriers would be likely to affect the implementation of the online peer tutoring intervention.

The process evaluation identified three main barriers to implementation; technical, logistical and finances. These finding will be used to inform any future studies involved in the implementation of online peer tuition interventions in schools.

Pupil and teacher views were mostly positive towards the perceived impact of the intervention and attendance in the lessons was high. Overall, the implementation of an online cross-age peer tuition programme between primary and secondary schools was found to be feasible, but with a number of technical, logistical and financial constraints.

5.8.5 Summary

The pilot efficacy study has demonstrated that it is possible to conduct small scale aggregated trials investigating the effectiveness of online cross-age peer tuition group sizes across the transition boundary between primary and secondary schools. The current chapter has addressed the research questions outlined in table 19, involving a data collection period of one academic year and 11 primary and secondary schools.

Table 19: Overview of research questions, methods, data and analysis for Study One.

Research question	Method	Data	Analysis	Conclusion
Is the implementation of an online cross-age peer tuition intervention feasible between primary and secondary schools as an aggregated trial?	Interviews, focus groups and lesson observations	Transcriptions of interviews and focus groups. Lesson observations.	Analysed using an inductive approach to Thematic Analysis (NVIVO software)	Yes this is feasible.
Is online cross-age peer tuition a feasible intervention strategy between schools as part of an aggregated trial?	Interviews, focus groups and lesson observations	Transcriptions of interviews and focus groups. Lesson observations.	Analysed using an inductive approach to Thematic Analysis (NVIVO software)	Yes this is feasible.
Are the instruments used appropriate for the ability of students involved in the study?	Pre-assessment	Pre-assessment data	Cronbach's alpha and Rasch analysis	Yes they are appropriate
Can the intervention maintain fidelity if delivered to schools in other locations in the UK?	Process evaluation (observation, interviews and focus groups)	Lesson observation and fidelity score	Comparison of the fidelity scores and lesson observation analysis	Unable to test due to school drop out.

Does one-to-two online cross-age peer-tutoring produce equivalent effects in mathematics attainment[11] compared with online one-to-one tuition?	Randomised Controlled Trial	Pre-test, post-test, dosage of intervention.	Multi-level model (linear mixed effects fit by REML)	Insufficient sample size to generate a conclusion.
Does one-to-many online cross-age peer-tutoring produce equivalent effects in mathematics attainment[2] as compared with one-to-one tuition?	Randomised Controlled Trial	Pre-test, post-test, dosage of intervention.	Multi-level model (linear mixed effects fit by REML)	Insufficient sample size to generate a conclusion.

The next chapter of the thesis will move onto the second primary research study to develop a feasibility study to create prospective cumulative meta-analysis based on aggregated small scale randomised controlled trials to evaluate the effectiveness of a TUTE mathematics intervention (Study Two).

[11] Mathematics attainment is defined as mastery level concepts in KS2 curriculum for Algebra.

Chapter 6: A feasibility study to develop a prospective cumulative meta-analysis based on aggregated small scale randomised controlled trials to evaluate the effectiveness of a TUTE mathematics intervention (Study 2)

TUTE Education was founded in 2011 with the aim of providing a social education for all children using small group learning with qualified teachers to make tuition affordable. TUTE has since developed into a leading online provider of small group lessons, used by over 420 UK schools and has delivered over 220,000 pupil places. TUTE currently provides 32 subjects taught from key stage 2 to key stage 5; however the majority of the interventions delivered are in the core subjects of English, Mathematics and Science. The lessons are live and not pre-recorded. The teachers are fully qualified and have a minimum of 3 years classroom experience ensuring the quality of the teaching is to a high standard.

TUTE has recently launched the UK's first Virtual Academy to help support schools to help plug the gaps in provision or to extend capacity, thereby delivering schools a flexible, on demand solution to help support teaching and learning. The evaluation is designed as an effectiveness study as the purpose is to evaluate the impact of TUTE when implemented in schools under normal school environments, rather than artificial conditions of a trial.

6.1 The intervention and rationale for conducting the research

6.1.1 Intervention

The TUTE intervention is delivered by UK qualified teachers from the virtual academy based in Wrexham. The intervention involved 5 progressive lessons on the topic of fractions. The lessons last for 45 minutes, with online sessions available throughout the school day, providing teachers the flexibility to engage with the intervention. These are delivered using the TUTE virtual classroom using headsets for the pupils to communicate with their teacher. The intervention programme was delivered over 5 weeks (outside of regular mathematics lessons) and the pupils randomised into the control 'business as usual' group will receive their usual numeracy lessons in their schools.

TUTE teachers were available from 9am – 4.30pm, Monday to Friday with the aim of enabling participating schools to deliver the intervention at a convenient time and with minimal disruption to the Year 7 pupils timetable. Three schools delivered the lesson at the same time every week, with two

schools opting for a rolling timetable so that the intervention did not affect the same non mathematics lessons each week.

TUTE provided each of the schools with usernames and passwords for the TUTE online learning platform, with the lesson accessible on the day of the lesson at the time booked by a link in the pupil timetable. Schools were required to have access to computers and headsets with microphones. Three schools were not equipped with appropriate headsets so these were provided by Wayne Harrison from Durham University as part of the online peer tuition project.

Prior to the start of the programme, TUTE conducted IT tests with the relevant IT technicians in each school, to ensure the TUTE software was accessible through school firewall systems. One school used two firewall settings, therefore in order for the TUTE platform to work the IT technician was required to open these 10 minutes before a lesson.

All pupils involved completed the pre-assessment and TUTE developed a learning programme for each school based around the topic of fractions. Figure 2.10 shows an example of a school learning programme.

tutè Learning Programme

Organisation and UID:		Subject:	Maths
Specific focus provided:	Tute effectiveness study with Durham University for 5 weeks.		
Key Stage and Year:	KS3 Yr7.6	Number of students:	12
Number of lessons:	5	Start date:	08.06.16

	Scheduled date:	Learning objective / focus:
1	08/06/2016	Why do we need to simplify fractions?
2	15/06/2016	What steps do we need to take to put fractions in order?
3	22/06/2016	How do we add and subtract fractions?
4	29/06/2016	What are the key differences when adding and subtracting fractions with different denominators?
5	06/07/2016	What are the similarities between multiplying and dividing fractions?

Additional information:

Please note:

Learning evaluations: Students' progress in all lessons are evaluated by the Tute teacher and will be presented to you in one evaluation per week. You will receive this learning evaluation within two days of the last lesson for that week.

Cancellation notice: 48 hours' notice is required to postpone or cancel a lesson. Lessons cancelled within 48 hours of the scheduled time will be debited.

Non-attendance: If the group does not attend a lesson and if the lesson has not been rescheduled, the content will be taught at the next scheduled time.

Figure 22: Example of a school learning programme

Each week the pupils logged into the TUTE lesson 5 minutes prior to the start time with the TUTE lessons scheduled for 45 minutes. However, one school reduced the lesson time to 30 minutes at the start of the intervention due to school logistical issues. Further details are included in the process evaluation.

Each lesson was delivered by the same TUTE teacher for a school partnership, with the tutor completing an evaluation of the lesson after each lesson. The lesson evaluations allowed the attendance to be monitored for each of the TUTE lessons delivered, enabling a sensitivity analysis to be conducted in the analysis.

Each school followed the same lessons, with the exception of the school with the reduced time of 30 minutes due to the weaker ability of the students and the time restriction.

Online learning platform

The same online learning platform outlined in Study One has been used for the research.

6.1.2 Rationale

The EEF has recently funded a trial to evaluate the effectiveness of one to one online mathematics tuition with Nesta and Third Space Learning (Torgerson et al., 2016). The intervention is delivered via online tutors trained in the UK National Curriculum and based in India. The trial was funded to provide tutoring via schools, making it available to disadvantaged pupils. The online delivery model is less well-developed and evidenced, so a trial was designed to test if remote tutoring is effective at improving academic results (Higgins et al., 2015).

The current study is designed to evaluate the effectiveness of online small group learning[12] in Mathematics compared to a business as usual control. TUTE provides online small group tuition in the UK and data collected by the company indicates that on average, pupil's attainment increases by 10% after a TUTE course[13]. However, case studies have numerous threats to validity involved in a single-group study: history threats, maturation threats, testing threats, mortality threats, instrumentation

[12] Online small group learning involves 1 qualified teacher and 12 pupils on the TUTE virtual platform.
[13] TUTE has three case studies available on their website from Pingle School, St Lawerence Academy and Manchester Communications academy (http://tute.com/how-it-works/#ourresults).

threats and regression threats. Therefore a robust evaluation is required to measure the impact of online small group learning.

Previous research acknowledges there have been few rigorous UK studies investigating the effectiveness of small group learning. Previous research by Elbaum et al., (2000), Slavin et al., (2011) and a recent evaluation by Torgerson et al. (2014) have found a positive effect size for small group tuition (weighted mean effect size[14] of 0.34). In terms of technology, the EEF toolkit shows that studies demonstrate moderate learning gains with a weighted mean effect size[4] of 0.28 (Higgins, 2015).

Unlike the previous trial evaluating one to one online tuition, the current study focuses on online small group learning, and adds to the evidence-base interventions that improve Mathematics in secondary school children. The recent trial conducted by Torgerson et al. (2016), used online tutors based in India to teach mathematics to primary school pupils in the UK. However, the trial found no evidence that the intervention had an impact on Key Stage 2 maths, compared with 'business as usual' teaching in Year 6. The online maths versus the control group reported an effect size (ES) of -0.03 (-0.35 to 0.28), with FSM affordable maths versus control ES -0.08 (-1.23 to 0.74). The study suggested further research to explore online small group learning as this might be a more efficient and cost-effective method of tuition for schools.

6.2 Methodology

6.2.1 Research questions

The evaluation aimed to investigate the effectiveness of a TUTE mathematics intervention on the mathematical skills of participating Year 7 pupils struggling with maths compared with Year 7 pupils continuing with 'business as usual' lessons.

The primary research question was:

> What is the effectiveness of a small group online mathematics programmes compared with "business as usual" on the maths skills of participating children?

[14] Cohen's d effect size used.

Secondary objectives included:

1) Does dosage impact on the effect of the intervention?
2) Is there a differential impact on pupil premium students[15]?
3) Is it feasible to implement a prospective cumulative meta-analysis for the aggregation of small scale randomised controlled trials in schools?

6.2.2 Trial Design

This trial was a pragmatic randomised controlled trial (RCT). A pragmatic design was chosen to reflect how the TUTE intervention is run in normal circumstances in schools. Consequently, teachers were given the freedom to select pupils for the study as they would do in normal teaching practice when deploying TUTE as an intervention. The trial involved 5 schools with 120 Year 7 pupils block randomised based on pre-test performance into either the intervention or control group, controlling for any unknown hidden biases when conducting the final analyses. The design used blocked randomisation using Alfiesoft data as pre-test ensuring the control and intervention groups are balanced for mathematics ability[16]. The trial is designed as an aggregated trial, as it was not feasible to conduct a large scale RCT with sufficient power. The aggregated approach allows the trial to be repeated with a minimum of two further cohorts required over the next academic year.

This is a waiting list design. This design is ethical as it does not deprive anyone of the treatment. It is less likely to demoralise schools / pupils in the control group than had there not been a waiting list. If the intervention is successful, the pupils will receive the intervention in the next academic year.

In order to illustrate the research design, the following notations from Campbell and Stanley (1963) are used:

X = TUTE intervention
O = Alfiesoft mathematics assessment
R = Randomisation at pupil level

[15] Pupil Premium children are identified as coming from disadvantaged backgrounds, such as low income families, looked after care or children from service families.
[16] Mathematics ability is defined by the topic assessed on Alfiesoft (fractions).

O R X O

O R O X

The main evaluative strength of the blocked RCTs is that each group is well balanced in all characteristics, with a very small chance of imbalance. Attrition prevents the full intention to treat analysis being carried out and can introduce bias (Dumville et al., 2006). A CONSORT flow diagram is included in the process evaluation in Chapter 7 to illustrate pupil opt-outs and baseline characteristics of participants lost to follow-up and those included in the analysis will be reported separately.

6.2.3 Eligibility criteria

Prior to randomisation, each participating secondary school identified 24 pupils who were eligible to participate in the trial. As part of reforms to the National Curriculum, the system of 'levels' previously used to report children's attainment and progress has been removed from September 2014 (Department of Education, 2014). In many educational trials, inclusion and exclusion criteria used reference levels to determine eligibility. A relative criterion, depending on each individual school assessment – "above expected", "expected" and "below expected" progress will be used in this study. Consistency between individual schools will be a limitation.

The following inclusion criteria used in the study are:

School inclusion criteria: Secondary schools are eligible to take part in the trial if they agree to all trial procedures including provision of pupil data, informing parents, randomisation and implementation of the intervention[17].

Pupil inclusion criteria: Pupils eligible for the intervention will be either "below expected" or "expected" for progress in Mathematics by the end of Year 7, based on teacher assessments.

6.2.4 Randomisation

Randomisation was conducted at pupil level in each school, carried out by an independent individual (independent teacher). Randomisation was blinded to prevent selection bias as a recent systematic review of trials found studies using non-blinded assessors generated more optimistic effect estimates than blinded assessors (Hrobjartsson et al., 2012). The 24 pupils were ranked in order of performance

[17] Appendix 5 includes the 'Agreement to Participate Form'

using Alfiesoft mathematics assessment pre-test (1 equating to best and 24 the least) and paired on ability. A coin will be used to select one of each successive pair of classes starting with 1&2 (then 3&4, etc.) forming the control and intervention group. This is a limitation and not the optimum method for randomisation. Through blocking the performance ability as indicated by the Alfiesoft pre-assessment on mathematics performance this will result in an equal distribution across the groups.

As previously identified in the primary Study One, only after the trial had commenced had the issue relating to potential bias through this form of randomisation had been acknowledged. In future, online randomisation would be used with the process fully recorded.

6.2.5 Outcome measures

Alfiesoft online assessments was the main test used to determine mathematics outcomes. The online assessments were automated marking, reducing the possibility of bias in the trial. The software is developed by AQA and used across many schools in the UK.

Primary outcomes

The primary outcome was the attainment score on the fraction assessment created using the Alfiesoft system. The assessment consisted of 19 questions and 41 marks. The purpose of the assessment was to assess the student's ability to: the use of fraction notation, understand simple equivalent fractions and simplify, calculate a given fraction of a given quality, understand and use fractions as multiplicative inverses and use efficient methods to calculate with fractions. In the assessment, a higher score represented higher attainment.

Delivery of outcomes

The researcher delivered the online assessments on an agreed date prior to the start of the intervention and in the final week of the planned online lesson. The assessments were completed under 'exam conditions' with all pupils (control and intervention) participating at the same time.

6.2.6 Sample size

The focus of this trial is on pupils who are performing at "below" or "expected" progress, therefore the sample size was based on this subgroup of children and limited to 12 pupil places in the TUTE virtual classroom. The initial cohort for the cumulative meta-analysis involves 5 schools, with 24 pupils

selected from Year 7. The total sample will be 120 pupils, 60 receiving the intervention and 60 continuing as business as usual. A recent EEF evaluation (Torgerson et al., 2014) found a positive effect size for small group tuition (weighted mean effect size[18] of 0.34). The aggregated study is powered to have an 80% chance of detecting an effect size of 0.34, with a sample of 360 pupils. However, through the prospective nature of a cumulative meta-analysis, repeating the trial and aggregating the results would allow for a more robust analysis of effectiveness.

The intra-class correlation rho = 0.15 accounts for the variance between schools, rho = 0.15 is selected as this is the recommended intra-class correlation for achievement in Mathematics (Tymms et al., 2011).

6.2.7 Recruitment

The schools initially targeted for participation in this trial were based in the North East of England, due to logistical reasons as this trial ran simultaneously with cohort 3 of the primary Study One. However, one school did not participate so a school from Derbyshire, already using TUTE was recruited as the 5th school. At the time of the study, TUTE did not have any schools using TUTE teacher interventions in the region, consequently schools participating in the online peer tuition project at Durham University were used as IT for the TUTE platform had already been tested. All participating schools provided parents with information and the opportunity to opt-out at any stage of the research. The next two cohorts of schools in the aggregated trial were originally planned to be recruited through TUTE schools, however the study was stopped after the first cohort had been completed.

Schools provided a list of all participating Year 7 pupils (2016/17) to Durham University identified by their name, gender and Free School Meal (FSM) status.

6.2.8 Data management

Data collection methods

A detailed project plan is included for the first cohort in appendix 5, including the schedules for the collection of baseline data, pre-assessments and post-assessments, lesson observations, post-intervention focus groups and teacher interviews.

[18] Effect size used Cohen's d.

Study instruments

The primary assessment used in this study involved the Alfiesoft online assessment. A serious limitation of the study is that the assessment is not a standardised assessment, as the questions selected for the assessment were selected by a teacher at TUTE. Ginsburg & Smith (2016) highlight this connection as a threat to the integrity of the study and this limitation will be discussed in further detail later in this thesis.

Trial diagram

The trial diagram outlines the pilot study for cohort 1 and a full CONSORT flow diagram is included in the process evaluation in Chapter 7 of this thesis.

Figure 23: Trial diagram for study 2 pilot cohort 1

Data Retention

The primary plan for the retention of participants involved the regular weekly contact with each of the participating schools through the lesson observations. The advantage of the convenience sampling used in this study, even though this provides a limitation, enabled the established relationships with schools to ease data retention. For instance, if pupils were absent for the pre-assessment the school conducted the assessment under exam conditions without the need for the researcher to be present. In the instance of a pupil leaving the school, it would not be possible to complete the assessment.

6.2.9 Statistical methods

6.2.9.1 Analysis

Analysis used intention to treat, therefore all pupils will be analysed in the group they were randomly allocated regardless of whether or not they received the intervention. Analyses were conducted using R statistical software, using a package developed by Durham University which is commonly used in EEF evaluations. Effect sizes are presented relating to the analyses alongside 95% confidence intervals. In this thesis the effect size is defined as:

$$g = \frac{\bar{x}_1 - \bar{x}_2}{s^*}$$

where the pooled standard deviation *s** is computed as:

$$s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

6.2.9.2 Baseline data

Pupil baseline data for the intervention and control groups are presented by trial arm to assess balance. Pupil level characteristics are summarised by trial arm for both as randomised and included in the primary analysis. Information on the lessons from which the pupils were withdrawn is also summarised. No formal statistical testing to assess balance was conducted.

6.2.9.3 Primary analysis

The primary objective of this study was to investigate the effectiveness of the TUTE intervention on the mathematics skills of the pupils in the intervention group. The primary measure was the Alfiesoft mathematics scores and this was analysed using a multi-level model. In the model, schools were random and the pre-test and post-test assigned as fixed effects.

The primary outcome measure is mathematics attainment and this was marked blind by the Alfiesoft online assessment.

6.2.9.4 Analysis – Secondary

A secondary analysis using the data from the intervention group using a multi-level was completed to determine the impact of the intervention on pupil premium funded students. An analysis on the dosage and behaviour for learning was also included.

6.2.9.5 Cumulative meta-analysis

Data had originally been scheduled to be collected in two additional cohorts in November 2016 and March 2017. Unfortunately the learning company TUTE decided not to continue with using evidence from randomised controlled trials and the study was unable to be completed. The study involved the cost of the intervention to be delivered in each of the schools and as the trial involved two further cohorts the decision was made to stop the trial.

6.3 Implementation and process evaluations

The process evaluation had two main purposes. Firstly, it assessed the fidelity of the delivery of the TUTE online intervention, as this is a fundamental prerequisite for the intervention. Secondly, it addressed the feasibility of conducting an aggregated trial for the effectiveness of an online small group tuition intervention.

6.3.1 Research questions

The process evaluation questions are:

- How feasible is it to evaluate the effectiveness of an online intervention using an aggregated trial methodology?
- How feasible and acceptable do teachers feel it is for using online small group tuition as a numeracy catch-up intervention?
- What are the barriers to the successful implementation of online small group tuition? How can these be minimised in future aggregated trials?
- What are the views of the intervention, of the peer tutors, tutees and teachers?
- What are the staff perceptions of the current and possibly sustained impact of the intervention on children's educational attainment?
- What are their suggestions for change if the intervention was to be more widely implemented?

6.3.2 Methods

The methodological considerations for choosing observation, focus groups and semi-structured interviews are outlined in Chapter 4 (p.91). As the intervention was delivered in schools already using the TUTE online learning platform for the online cross-age peer tuition, no prior testing of the online platform was required. The process evaluation consisted of two stages. Firstly, observations were completed in the North East schools for each of the lessons delivered. It was not feasible to observe the lessons from the school recruited from the midlands, however the TUTE lesson feedback provided evidence that the lessons had been completed. A fidelity measure was recorded on a scale of 0 to 9 for each lesson observed. In addition to the lesson observations, the teacher delivering the intervention also completed a lesson feedback form. The second stage of the process evaluation

involved conducting focus groups with the pupils participating in the intervention (2 schools selected randomly due to time constraints) and semi-structured interviews with two teachers from the schools randomly selected. These are outlined in the detailed project plan in appendix 6.

6.3.3 Fidelity

Lesson observations were completed for all the TUTE lessons delivered in schools located in the North East of England. Fidelity was assessed through a researcher observation measure and a lesson evaluation completed by the TUTE teacher delivering the intervention.

The researcher observation measure consisted of three component scores relating to (a) how well the technology worked (b) the students attitude to learning (c) the students behaviour for learning. For each of these components was rated between 1 and 3, with 1 corresponding to 'poor', 2 corresponding to 'fair' and 3 corresponding to 'excellent'. Therefore, each online tuition lesson observed by the researcher had a fidelity score which ranged from 3 to 9. Lower scores corresponded with lower fidelity.

As part of the TUTE learning programme, each teacher delivering a lesson on the platform is required to complete a lesson evaluation. Figure 24 provides an example of the TUTE teacher evaluation form.

Organisation/UID	School 2	No of pupils	11/12
Subject	Maths	Date and time	06.07.16 Wednesday 10:20
Key Stage / Year	KS3 Yr7	Teacher	Teacher B
Learning objective:	How do we multiply and divide fractions?		
Learning outcomes:	• To multiply fractions • To divide fractions • To calculate a fraction of a number		

Pupil	Achievement	Effort	Behaviour for learning	Comments/Suggestion for progress:
	4	4	5	
	3	4	5	To flip the second fraction when dividing.
	4	4	5	
	3	4	5	To flip the second fraction when dividing.
	4	4	5	
	3	4	5	
	4	4	5	
				Absent
	4	4	5	
	3	4	5	To flip the second fraction when dividing.
	4	4	5	
	4	4	5	

Additional comments:
Al students took an active part in the lesson where they could. A, B and C had problems with screens freezing and sound issues and were unable to follow parts of the lesson. They all need to practise recognising fractions that can be simplified. Over the past 5 weeks these students have been a pleasure to teach and I wish them all the best of luck in the future.

Figure 24: Lesson evaluation form

6.4 Ethics and dissemination

6.4.1 Ethical approval

Approval from Durham University School of Education Ethics Committee was sought prior to the study starting and granted on 4/5/2016.

Schools informed parents of the pupils participating about the study with a parental information letter and an opt-out letter will allow parents the opportunity to withdraw their child's data. Appendix 6 includes the parental letter and a copy of the opt-out ethics form.

6.4.2 Protocol amendments

Amendments to the protocol are:

1. School 3 selected a group of pupils from the nurture group rather than the targeted approach used by the other 4 schools. Therefore the overall ability of these students were lower and this is discussed in the process evaluation.

2. School 3 changed the time from 45 minutes to 30 minutes, in order to allow the intervention to be run over the lunchtime.
3. TUTE decided against running the second and third cohorts so the study was terminated after the initial pilot. Therefore no aggregation of the data is possible.

6.4.3 Confidentiality

The data was stored on a password protected computer and backed up on eth secure university server. All student, teacher and school identifications are replaced with anonymous codes to ensure no identification is possible.

6.4.4 Dissemination

The dissemination plan involves creating a technical brief for TUTE with a detailed process evaluation to support the development of the online lessons. An impact report was produced for each of the participating schools in the study. The findings of the aggregated trial would be written as an academic journal article and submitted for peer review and publication, if the full trial had been completed.

6.5 Methodology Summary – Study 2

The second study in this thesis uses a randomised controlled trial design to test the effectiveness of small group online mathematics teaching. Due to the withdrawal of the support for the use of randomised controlled trials from senior leaders as the educational company TUTE, only one cohort of schools (5 schools) out of the proposed sample of 15 were involved. Further discussions around the issues in evaluating educational interventions developed in business will be explored in detail later in this thesis.

Table 20 provides an overview of the research questions involved in study 2, with a description of the method, data collected and analysis. The full analysis for the impact and process evaluation for study 1 can be found in Chapter 7.

Table 20: Overview of study 2

Research question	Method	Data	Analysis
What is the effectiveness of small group online mathematics programmes compared with "business as usual" on the maths skills of participating children?	Randomised Controlled Trial	Pre-test, post-test ALFIE-soft data	Multi-level model (linear mixed effects fit by REML)
Does the dosage analysis impact on the effect of the intervention?	Randomised Controlled Trial	Pre-test, post-test, dosage of intervention.	CACE analysis
Is there a differential impact on pupil premium students?	Randomised Controlled Trial	Pre-test, post-test.	Multi-level model (linear mixed effects fit by REML)
Is it feasible to implement a prospective cumulative meta-analysis for the aggregation of small scale randomised controlled trials in schools?	Process evaluation (observation, interviews and focus groups)	Lesson observation, TUTE teacher feedback and fidelity score	Comparison of the fidelity scores and lesson observation analysis

Chapter 7: Impact and process evaluation for study 2 investigating the effectiveness of a TUTE mathematics intervention

The focus of this chapter is to present the findings from Study 2 methodology outlined in Chapter 6 investigating the effectiveness of an online small group tuition for a TUTE mathematics intervention. The first section outlines the impact evaluation and the second section presents the process evaluation. The initial design of the aggregated RCT planned three stage data collection period to evaluate the effectiveness of online TUTE lessons delivered by qualified teachers in small groups. Unfortunately, only one stage of the data collection was permissible. Therefore, it is important to note that the conclusions from this report are not intended to be generalizable.

The intervention aims to help improve pupils' maths skills while they are in their first year at secondary school (Year 7), specifically targeting the maths skills of pupils who are not making expected progress. In the evaluation, the control condition is a 'business as usual' with a waitlist as the pupils not completing the intervention will be offered this in the following year. TUTE was responsible for the design of the mathematics lessons and delivering the planned mathematics intervention on fractions to Year 7 pupils in 5 secondary schools. An independent online assessment was used as the primary assessment outcome, delivered as a pre- and post-assessment in each of the pilot schools.

In advance of the chapter commencing, it is useful to revisit the primary and secondary research questions. Table 21 provides an overview of the research questions, method, data and analysis used in the feasibility study and first planned wave of the small scale randomised controlled trials for the development of a prospective cumulative meta-analysis.

Table 21: Overview of Study 2 research questions, methods, data and analysis.

Research question	Method	Data	Analysis
What is the effectiveness of small group online mathematics programmes compared with "business as usual" on the maths skills of participating children?	Randomised Controlled Trial	Pre-test, post-test ALFIE-soft data	Multi-level model (linear mixed effects fit by REML)
Does the dosage analysis impact on the effect of the intervention?	Randomised Controlled Trial	Pre-test, post-test, dosage of intervention.	CACE analysis
Is there a differential impact on pupil premium students?	Randomised Controlled Trial	Pre-test, post-test.	Multi-level model (linear mixed effects fit by REML)
Is it feasible to implement a prospective cumulative meta-analysis for the aggregation of small scale randomised controlled trials in schools?	Process evaluation (observation, interviews and focus groups)	Lesson observation, TUTE teacher feedback and fidelity score	Comparison of the fidelity scores and lesson observation analysis

7.1 Logic model / theory of change

A logic model or theory of change (Table 22) reflects the views of the management of TUTE on the necessary conditions (inputs and processes in the school) for their intervention to be successful. The primary purpose of the logic model is to provide the underlying assumptions about how the intervention will enable the expected outcomes to be achieved (Clapham et al., 2017). The process

evaluation will provide evidence on whether these conditions were successfully achieved in the implementation of the intervention[19].

Table 22: Logic model

Inputs	Processes in school	Change mechanism	Pupil intermediate impacts	Outcomes
• Sufficient computers with access to the internet and headsets with microphones • Good classroom environment with sufficient space between students on the computers • Online mathematics teacher from Tute • Online lesson resource created by Tute • In school supervision (TA or teacher)	• Online 45 minute maths lesson for 5 weeks • Online lessons in addition to mathematics lessons in school • TA or teacher to supervise log in and behaviour during the lesson	**Education** • Children learn mathematics with a qualified teacher in an online synchronous learning environment • Tute teacher uses engaging online lesson materials • Additional learning support to complement classroom practice **Online delivery** • Online delivery to engage students learn mathematics via a different medium to traditional classroom learning • Small group learning* allows more opportunities for students to ask questions and receive support	• Improved concentration and engagement in mathematics • Improved confidence and self-esteem • Improved mathematical reasoning skills • Improved communication skills	• Increase in mathematics attainment on Afliesoft assessments

[19] The intervention involves mall group learning in TUTElessons involve 12 students and 1 qualified teacher. This is larger than traditional small group learning (4-5 students).

7.2 Impact evaluation

7.2.1 Participants

Each school participating in the trial was asked to identify 24 pupils in year who were working below expected progress, ideally in the middle sets in Year 7 (sets 3,4,5). Four schools adhered to the guidance but as this was a pragmatic trial, one school (school 3) decided to use bottom set pupils (set 7) in the trial. The decision to include pupils outside of the original eligibility criteria threatens the validity of the study, as the content had originally been designed for pupils for a higher ability. This potentially introduces a negative bias for evaluating the impact of the intervention. In total, 5 schools recruited 24 pupils with a total sample size n=120. The intervention group n=60 would receive 5 TUTE lessons and the control n=60 would continue as "business as usual".

As the CONSORT diagram in figure 1 shows, 2 pupils in the intervention group were withdrawn from the trial before, one as a pupil did not wish to participate and another pupil moved schools. Therefore, 58 pupils received the intervention and 60 pupils were part of the control. During the data collection stage, one pupil was ill for the last week of term in the intervention group (n=57) and three pupils were unable to complete the post-test. Reasons included two exclusions and one absence due to illness, with the control group n=57.

Therefore, attrition in the trial was 5% and balanced between the control and the intervention arms in the trial.

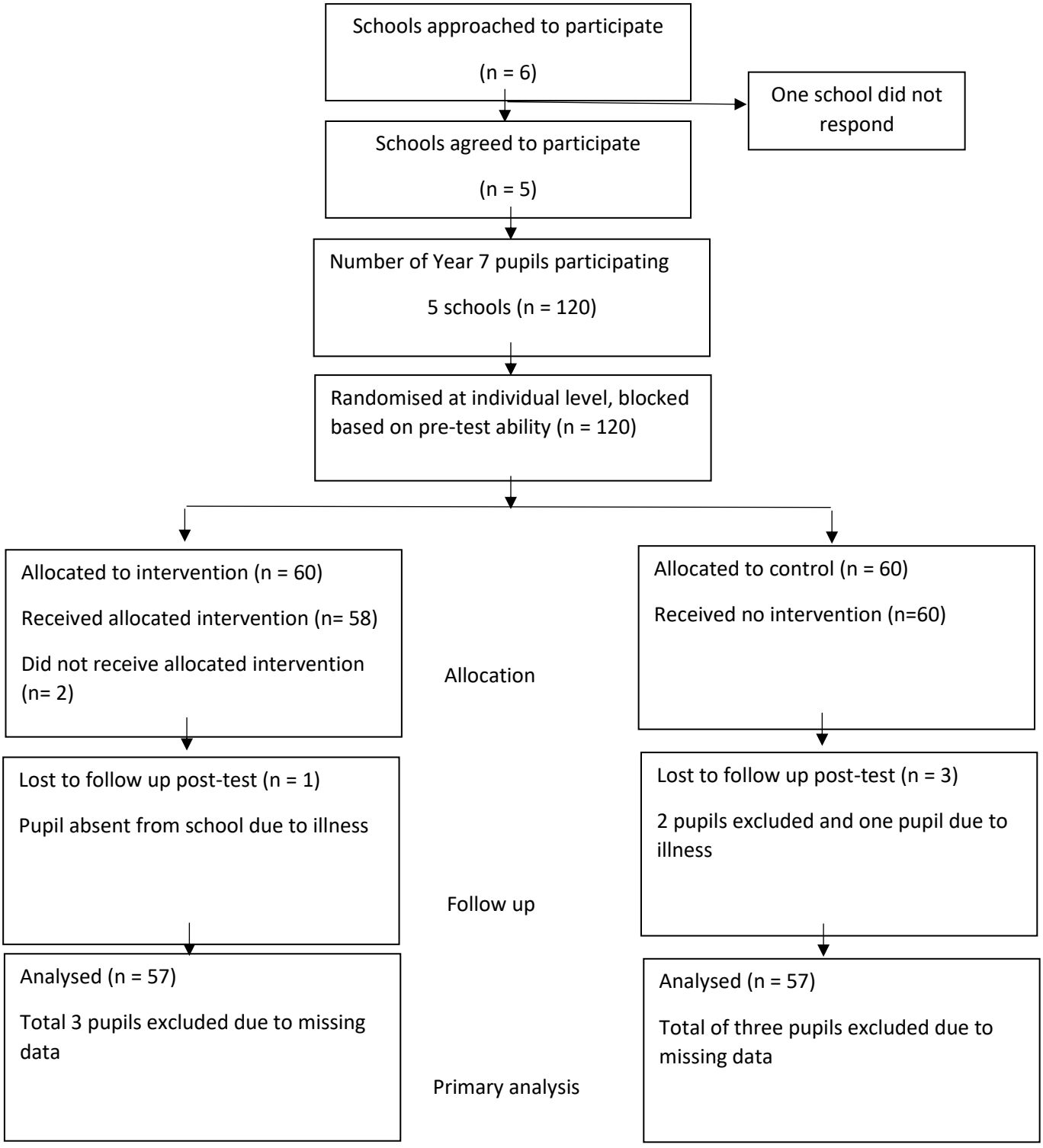
Characteristics of the 114 pupils included in the analysis are reported in Table 23 with data available on the pupils missing.

Table 23: Pupil characteristics

	Control	Intervention	Missing
sample	57	57	6 (pre-test data on 4)
Pre-test score	10.87 (SD 8.99)	10.96 (SD 8.65).	6.00 (SD 2.16)
Male	35 (61%)	32 (56%)	1 (25%)
FSM	16 (28%)	20 (35%)	2 (33%)
EAL	3.5%	1.54%	0%

7.2.2 CONSORT Flow Diagram

Figure 25: Study Two CONSORT flow diagram Cohort 1



7.2.3 School characteristics and summary of each school context

Four of the five schools are located in the North East of England with one school located in Derbyshire. The randomisation of the pupils occurred at individual level, therefore no schools were involved as a control so no detailed analysis has been included. Table 24 outlines the key characteristics of the schools.

Table 24: Characteristics of the schools

School	Number pupils	KS2	FSM	EAL	IDACI
0	871	4.5	32.8%	1.6%	17.3%
1	527	4.7	27.1%	1.5%	24.7%
2	852	4.8	17.0%	3.2%	13.0%
3	1006	4.5	52.5%	1.1%	78.3%
4	813	4.6	49.9%	1.2%	57.3%

7.3 Outcomes and Analysis

As outlined in the statistical analysis plan (Appendix 6) the analyses of primary and secondary outcomes were undertaken on an 'intention to treat basis', meaning that all those allocated to treatment and control conditions in the randomisation are included in the final analysis, even if they were to drop out of the treatment (Torgerson et al., 2005).

7.3.1 Descriptive analysis

Schools were asked at baseline when they envisaged implementing the TUTE online intervention: during a school lesson (not mathematics), at break time or after school. Four schools (80%) selected during a school lesson, one school (20%) selected at breaktime. The lunchtime break was not recommended to the school as the intervention time would need to be reduced to 30 minutes instead of the usual 45 minutes. Unfortunately, the only time the intervention could be delivered was during lunchtime so the lessons were adapted to fit the reduced time allocation.

The most commonly selected lesson objectives were 'Why do we need to simplify fractions?' (100%), 'What steps do we need to take to put fractions in order?' (100%) and 'How do we add and subtract fractions?' (100%). The remaining two lesson slots varied between schools with 'What are the key differences when adding and subtracting fractions?' (60%) and 'What are the similarities between multiplying and dividing fractions?' (60%).

At the end of each lesson the teacher delivering the TUTE intervention completed a lesson evaluation rating the achievement, effort and behaviour for learning for each of the learners from a scale of 1 to 5 (a score of 5 is excellent). A total of 25 TUTE lessons were completed with an average rating for achievement 3.71, effort 4.11 and behaviour for learning 4.15. Table 25 shows the average per school for the five lessons delivered in the programme.

Table 25: Fidelity in schools participating in the intervention

School	Achievement	Effort	Behaviour for learning
0	3.68	4.12	4.18
1	3.94	4.48	4.74
2	3.9	4.42	4.86
3	3.3	3.66	3.22
4	3.72	3.86	3.76

The raw data for the pre- and post-assessments are presented in Table 26 for the schools participating in the TUTE evaluation.

Table 26: Pre and Post assessment scores

	School				
	0	1	2	3	4
Intervention (n=57)	n= 11	n= 12	n=12	n= 11	n=11
Pre-test	5.58	11.00	24.58	5.73	6.73
SD	2.97	3.69	9.65	5.41	2.90
Post-test	6.00	13.66	27.50	4.55	8.3
SD	3.36	4.87	7.49	4.49	2.41
Control (n=57)	n= 10	n= 12	n= 11	n= 12	n= 12
Pre-test	5.83	10.83	25.72	5.92	6.50
SD	2.86	3.88	8.90	4.71	2.02
Post-test	6.5	9.67	25.82	6.27	7.42
SD	2.83	1.78	9.88	4.52	1.78

The mean pre-test score the control group (n=57) was 10.87 (SD 8.99) and the mean pre-test for the intervention group (n=57) was 10.96 (SD 8.65). The mean post test scores for the control group 10.89 (SD 8.84) and the intervention group 12.73 (SD 9.81). The pre-test was included in the model

as a raw score at the pupil level. This will account for part of the variance explained at the pupil and school level (Dwan et al., 2008).

7.3.2 Intra-cluster coefficients and correlation (ICC)

The intra-cluster correlation coefficient was estimated using the data from the assessments to be 0.28.

7.3.3 Multi-level model

The primary outcome analysis was conducted using a linear mixed effects model fit by REML (Restricted Maximum Likelihood). In the model, the school was random effects with pre- and post-tests as fixed. The school is random as it is assumed the effect varies randomly within the population of the organisation. Pre- and Post-tests are fixed as it is assumed to be measured without measurement error and the variable used in the study contains all or most of the variables values in the population.

The following assumptions in the model for multiple group comparisons are:

5. Independence of observations within groups
6. Independence of observations between groups
7. Normal distributions of observations
8. The groups share a common variance

The rationale for using a Multi-Level Modelling (MLM) is the assumption of independence within schools is most likely to be violated. When considering the normality assumption, normal t-based procedures are robust to it, which means, even if the normality assumption is not met and sample is sufficiently large enough, the procedure would produce more or less the same results. Yet, they are

sensitive to the violation of equal variance assumption. For example, in smaller sample (n), it is more likely to reject the null hypothesis too often (false positive, type 1 error). In larger n, it is more likely to under-reject the null when the alternative is true (type II error).

The total variability is used because those derived from between and within variances can be biased, representing the most conservative approach and least likely to result in false positives compared to using within or between variability (Kasim, ZhiMin & Higgins, 2015).

The effect size were calculated using Hedges g is +0.44 (95% CI – 0.3to 1.17). This is a good effect size and demonstrates that the children who received the TUTE intervention scored higher than the children who did not. However, an analysis of the effect size in the model at each individual school demonstrates large variation of the effect of the intervention in each school.

Table 27: Effect sizes by school

School	Effect size (Hedges' g)
0	-0.24
1	-0.01
2	0.74
3	-0.57
4	0.06

Possible explanations for the variance are linked to the behaviour, ability and attitude to learning of the pupils participating in the intervention. School 3 had a reduced lesson time and weaker ability students compared to the edibility criteria initially discussed with schools. This school used the lowest set Year 7 pupils compared to selecting pupils identified as below progress from middle set classes. The average pre-test score for School 3 was 5.83 (SD 4.94) compared to School 2 mean pre-test score 24.29 (SD 9.8). This indicates the pupils in school 2 were of a higher ability compared to the other participating schools, whilst demonstrating exemplary behaviour during the lesson observations. This

was confirmed with an email written by the researcher to the head teacher commending the attitude of the pupils to learning at the end of the project and before the post-test had been administrated.

7.3.4 Secondary analyses

Effect on attendance (CACE analysis)

A secondary analysis was conducted using the CACE from the R statistical package developed by a team at Durham University for the EEF (Kasim et al., 2017).

	Compliance	ES	LB	UB
1	P> 0	0.44	0.14	0.75
2	P> 10	0.44	0.14	0.75
3	P> 20	0.44	0.14	0.75
4	P> 30	0.44	0.14	0.75
5	P> 40	0.44	0.14	0.75
6	P> 50	0.44	0.14	0.75
7	P> 60	0.47	0.15	0.80
8	P> 70	0.47	0.15	0.80
9	P> 80	0.66	0.22	1.10
10	P> 90	0.66	0.22	1.10

The data shows compliance as a percentage of attendance with a predicted effect size for the intervention. However, this should be treated with caution as the length of the TUTE intervention in the trial was only 5 lessons. It is recommended that future trials, outside of this study are run over a longer time period.

The analysis deviated from the statistical analysis plan by not completing an analysis for the pupil premium subgroup analysis. This is due to the limited sample size of the study.

7.4 Process evaluation (Part 1)

The process evaluation had two main purposes. Firstly, it assessed the fidelity of the delivery of the TUTE online intervention, as this is a fundamental prerequisite for the intervention. Secondly, it addressed the feasibility of conducting an aggregated trial for the effectiveness of an online small group tuition intervention.

The next section outlines the key components of the process evaluation, providing context to the implementation of the intervention prior to addressing the process evaluation questions outlined in chapter 2.

7.4.1 Implementation

The delivery of the intervention mainly occurred in either specialised computer rooms (4 schools) or a standard classroom using laptops. Typically, pupils arrived 5 minutes before the start of the TUTE lesson for 18/25 lessons observed, to start up the computers and log into the TUTE online platform. A number of schools did not have the correct headsets with microphones, so these were provided by the researcher in advance of eth intervention starting in the schools.

7.4.2 Length and timings

The lessons were approximately 45 minutes long once the pupils had logged onto the computer and started the TUTE intervention lesson. The exception was one school, whereby the initial time had changed due to timetabling issues, so this was reduced to 30 minutes in tutorial time (school 3). One teacher commented that they thought the 45-minute lesson was too long for the year 7 pupils (school 0), as their concentration reduced after 30 minutes. The lesson observations also noted that off task behaviour such as pupils putting ear phones over their heads (not ears), attempting to distract other pupils' attention, changing the TUTE format on the screen (chat box on full screen rather than lesson)

and other off task behaviours increased towards the end of the lessons. It is possible that pupils with higher ability may be able to concentrate for 45 minutes but it did appear that this time was slightly too long for the ability of students involved in the trial. The times of the delivery varied, with only one school using tutorial time. A number of students in the control groups involved in the focus groups complained about missing either drama or PE. However, this is no different from pupils usually missing classes due to planned other interventions in the schools.

7.4.3 Target group

From the five schools involved in the trial, the researcher briefly discussed the intervention with the teachers supervising the TUTE lesson. Generally, the feedback was positive with only one teacher explaining that she did not think the programme worked well. In this school, logistical issues surrounding the delivery occurred with the intervention delivered in tutorial time. However, it should be noted that the teacher in question struggled to supervise the intervention due to school commitments with the researcher supervising two of the five lessons.

7.4.4 Technical set-up

Initial set up

The technical issues related to the set-up had fortunately been resolved in the parallel trial using the TUTE platform for the online cross-age peer tuition intervention earlier in the academic year. However, the TUTE programme requires IT testing using TUTE technical support and an IT technician in each school. Two schools encountered problems with the testing of TUTE, one school had two separate firewalls requiring IT support each lesson to disable these and the other school had issues with audio due to the software installed on computers. These issues were managed throughout the project and did not impact on the delivery of the lessons.

On-going technical issues

The initial logging on in a number of schools was stressful for the teachers and lengthy, in particular with weaker ability students. For instance, in the school with 30-minute lessons a number of pupils struggled to type in their username and password correctly. The issue involved pupils using the 'a' key on the keyboard instead of the @. In one school, the google chrome had been removed from a number of computers and this required the technician to install this again on the computers. Additionally, in two schools the issue regarding damage to computers created a problem. Often, chewing gum had been placed in the headset sockets or keys had been changed location on the keyboard, confusing the Year 7 pupils. As previously mentioned, the sound would often drop out due to software on the school pcs on one particular school so in order to avoid this, the teacher marked the computers that worked with stickers so that these were used each week. Finally, if laptops used the school wifi, the lessons would sometimes freeze and lose connection. This only occurred on a few occasions but it did disrupt the learning of the pupils involved, as it required them to log out and load up the lesson again. After speaking with TUTE IT technicians, it was advised to try to use computers connected directly to the internet.

7.4.5 Online delivery

The pupils generally enjoyed the TUTE lessons, with a number of pupils commenting after the lessons how much fun the lesson was and how it had helped them learn fractions. A difference in teaching style was noted from the lesson observations, with a number of pupils commenting on the fun aspect as the teacher used a verbal countdown for the pupils to answer questions in the chat box (School 2). This was in contrast to the other style whereby pupils could answer when they had worked out the answer. The lesson observations noted that the pupils without the countdown sometimes blatantly copied other students work. When I asked a pupil how they got the answer to a question, they responded:

"I just copy the answers of student X, as she is cleverer than me"

This was observed more frequently in the lessons without the countdown and competition element in the online lesson. Another common feedback from the pupils from across all the schools was the inability of the tutor to see when they were typing the chat box, with the tutor inadvertently interrupting them when they were thinking through an answer. A solution provided by a few students would the ability for the TUTE platform to be similar to Facebook messenger where you can see a person is typing.

The TUTE lessons observed were structured well, with opportunities for students to reflect on their learning and polls used by the teacher to assess the progress throughout the lessons. As the polls were not visible to other students, it was less likely that pupils copied the answers of fellow peers.

A potential distraction noted in the lesson observations was the ability of the pupils to change the layout of the TUTE lessons, regarding the size of the chat box, slide presentation and webcam. It was noted on a number of occasions pupils were clicking the minimise screen on the slides and increasing the size of the chat box. They were therefore unable to see the slides with the questions.

7.4.6 Behaviour and attitude

The lesson observations clearly showed a difference in the attitude to learning across the five schools participating in the trial. School 2 in the analysis had very few behaviour issues throughout the trial will pupils arriving early for the online lessons, engaging in the competitive lessons as the teacher used strategies to engage the pupils (count down, games, enthusiastic delivery). However, school 3 involved pupils of a lower ability compared to the other schools participating in the trial. For example, in the first session a number of students required support in logging into the computer and TUTE platform as they struggled with the @ symbol in the username. In this group of students, off task behaviour was common amongst the pupils with the lesson reduced to 30 minutes due to timetabling issues.

Therefore pupils missed the tutorial part of lunch rather than lesson time. School 0 was not observed during the trial but the teacher commented that the lesson time was too long for the ability of students, with attention decreasing after 30 minutes.

7.4.7 Human investment

The initial set up requires time from the IT technicians in schools, as the IT tests had already been checked with a previous project these were not an issue in this study. In order for TUTE to be delivered in schools, the pupils require supervision either from a teaching assistant or teacher. In three of the schools, teachers supervised the lesson and teaching assistants were deployed in the remaining schools. Ideally, the TUTE resource does not require teacher supervision. However, these schools have limited teaching assistant resources so teaching staff were used as supervision.

7.4.8 Support and commitment

The support to schools provided by TUTE was excellent in terms of the technical set up, creation of pupil accounts and arranging the lessons for delivery. One mathematics session in a school clashed with the England football match, with all the pupils in the school off timetable to watch this in the school hall. The teacher informed TUTE and the lesson was moved to the following week and the project extended by the seven days. Additionally, when technical issues did occur such as pupils becoming locked out due to typing an incorrect password 3 times, the account was unlocked within 5 minutes of the incident occurring.

Another example for the flexibility and response to school issues included the initial week with the school reducing the lessons to 30 minutes. As the school had involved weaker students in the trial, the TUTE teacher communicated with the teacher to change the structure of the lessons to make these

more accessible. This is common practice with the TUTE teachers responding to the needs of the pupils.

7.4.9 Fidelity

The intervention was implemented as intended, and fidelity to the trial conditions was high. The researcher observed 18/25 lessons delivered by TUTE, with a conservation with the member of staff after each session which was not observed (phone call or email). Four schools complied with the eligibility criteria and adhered to the full-time allocation for the TUTE lesson. One school reduced this to 30 minutes and included pupils with lower levels of mathematics ability. These were the students who struggled to log into the computer and the TUTE platform. The fidelity in school 0 is considered low and a possible explanation for the large negative effect size (-0.57).

The TUTE teacher learner reports provide supporting evidence for the high fidelity in schools 0, 1 and 2 and the lowest score from school 3. Appendix 12 outlines the fidelity measures in further detail.

Table 28: Fidelity reporting at school level

School	Achievement	Effort	Behaviour for learning
0	3.68	4.12	4.18
1	3.94	4.48	4.74
2	3.9	4.42	4.86
3	3.3	3.66	3.22
4	3.72	3.86	3.76

Discussions with all the schools indicated that the control groups were not being taught fractions at the time of the trial and all the pupils in the trial were not receiving additional mathematics interventions.

7.5 Process evaluation (Part 2)

The process evaluation consisted of three main methods of data collection. Firstly, a large proportion of lessons (80%) were observed by the evaluator with a fidelity score assigned to each lesson. Secondly, focus groups and interviews were conducted with two schools located in the North East with telephone conversation with the teacher in the school located in the Midlands. Thirdly, after each lesson the teachers at TUTE completed a lesson evaluation form rating the achievement, effort and behaviour for learning.

Semi-structured interviews were conducted with teachers in three schools and were recorded using an audio device. The focus groups and interviews were transcribed then imported into NVIVO software. The data was analysed using an inductive approach to a Thematic Analysis based on the framework by Braun & Clarke's (2006) 'guide' to the six phases of conducting this process. A limitation with the analysis was that only one person coded the data, therefore interrater reliability was not possible. However, a coding framework was developed to ensure the method was applied consistently.

The process evaluation research questions which are set in in Chapter 6 will now be revisited and answered using supporting evidence from the lesson observations, TUTE teacher feedback forms, pupil focus groups and teacher interviews.

7.5.1 How feasible is it to evaluate the effectiveness of an online intervention using an aggregated trial methodology?

The feasibility study has demonstrated the ability for a single researcher to work with schools to conduct a small-scale aggregated trial in school settings. Previous research (Gorard, Siddiqui, and See, 2016) has demonstrated the ability of teachers to conduct randomised controlled under the supervision of independent researchers in a single concurrent timescale. The purpose of the feasibility

study involved the aggregation of cohorts of the same trial to allow the data to form the evidence for a cumulative meta-analysis.

A major limitation of the study was the decision by the online company to withdraw from the study before the planned second and third cohorts. Consequently, the ability to aggregate the data was not possible with the current study.

7.5.2 How feasible and acceptable do teachers feel it is for using online small group tuition as a numeracy catch-up intervention?

Interviews with the three teachers were generally positive towards the online small group tuition intervention. A common theme from the interviews was the use of technology to engage the learners in mathematics. Teacher A commented:

> *"It is clear the students engaged as the usual off task chatter was minimal in all the lessons. As soon as the headsets were on, the room went silent and the pupils were listening to the teacher."*

Teacher B reinforced the engagement and lack of off task behaviour:

> *"The headsets appeared to help the pupils focus on the tasks and throughout the delivery of lessons I hardly heard any off-task behaviour. For these pupils, this is rare".*

When asked to clarify what the teacher meant by off-task behaviour, they responded:

> *"If you observe the pupils in a traditional mathematics class, within minutes of starting tasks you will hear conversations not related to maths. It is hard to get students to concentrate for 30 – 45 minutes without distractions".*

Another key theme emerging from the interviews was the group sizes of the learner to teacher ratio. All teachers highlighted that they thought the small group size of 12 was slightly too large for teaching online. Teacher B explained:

> *"The difficulty the online teacher has is that they have to engage 12 pupils in the online classroom as the same time. This is fine for explaining a concept, however pupils work at different paces and a number of times I noticed pupils copying answers from other pupils in the chat box".*

The teacher expanded on this further:

> *"If the group sizes were smaller, pupils could take turns answering questions or working in pairs for the teacher to question their understanding. If the technology could allow it, having pupils working in their own space would prevent pupils copying. The teacher could then select a pupil example to go through with the group."*

Teacher C explained that the small group size was actually very similar in size to a lower ability class size for her Year 9 maths set.

> *"I understand that for the business they are delivering the intervention using one teacher to twelve learners, probably for cost effectiveness for schools. Yet, I teach a year 9 low ability class with 14 pupils. In my opinion a small group would be four to five pupils".*

When asked about how feasible it is for schools to use online small group teaching as an intervention, all three teachers were generally positive about the approach of the intervention. A theme across the three teachers' responses focused on the lack of time after school for interventions with KS3 pupils.

Teacher C explained:

> *"Due to after school meetings and GCSE revision for Year 10 and 11 pupils, I physically do not have the time to deliver interventions to Key Stage 3 pupils. The ability to outsource this to another qualified teacher and have the intervention run in the library would allow these pupils to receive support".*

However, two of the three teachers explained their concerns over shrinking school budgets and the cost would not be able to be funded from with their department budgets. When asked if financial constraints would be an issue for deployment in their school, all three agreed that they would require significant funding from additional budgets such as the pupil premium.

7.5.3 *What are the barriers to the successful implementation of online small group tuition? How can these be minimised in future aggregated trials?*

The potential barriers to the successful implementation of online small group tuition can be classified into three main points:

4) Technical – A fundamental component to the success of the intervention is the quality of the IT infrastructure within the schools. IT infrastructure includes the quality of internet (bandwidth), firewall permissions, compatible computers and maintenance of devices in s school environment.
5) Logistical – The time required to deliver the lesson in school, access to computer rooms, headset availability and staffing requirements.
6) Financial – Sufficient budgets within schools to purchase the intervention.

In order to minimise the three main barriers to implementation for future aggregated trials involved in online small group teaching, each barrier can be reduced through the following:

Technical

The company has clear guidance for school IT technicians to clear the firewalls to allow the software to access the relevant ports for video and sound functionality. Due to the nature of the collaborative learning space software (flash-based component), two schools (40%) had issues bypassing the firewalls. These were managed for the pilot and did not affect the delivery of the intervention, however at scale a permanent solution such as migrating to HTML 5 software could reduce the firewall problems.

In actual lessons, the technical issues most frequently encountered involved the maintenance of the computers used for the intervention. Providing clear guidance to schools to use sound ports at the back of computers (less likely to be damaged) and to ensure computers had a mouse and keyboard

keys are not missing. Good practice observed in two schools involved teachers marking computers with stickers, ensuring learners used the same computers each week.

The platform required PC / laptops for the online classroom to work. If the solution worked on tablets, it is likely that the issues raised above will be minimised. However, working with pupils in schools will unfortunately undoubtedly cause issues relating to damage to equipment.

Logistical

The main logistical issue for schools involved locating a computer room during the school day which was not used for lessons. In school environments, it will not be possible to minimise this particular barrier to the often limited availability of computer classrooms. The second main barrier for logistics related to access to headsets with microphones, as this presented a challenge in four schools (80%). Providing headsets as part of the intervention will minimise the risk, however it is still possible for these to go missing once in a school environment.

Financial

The platform and lessons were provided at no cost to schools participating in the initial feasibility study, yet as a business model moving forward this would not be sustainable. In order to scale, independent funding for a trial from independent companies such as the EEF would need to be explored.

7.5.4 What are the views of the intervention of the learners and teachers?

Learner views

The focus groups with Year 7 pupils found generally positive views.

> *"I found it easier to learn with a headset on as I was not distracted; it has helped me learn maths."*

"The teacher was really good, they helped explain how to multiply fractions and I feel more confident now."

"I usually find fractions difficult and boring but it was great learning with a teacher in an online classroom."

Two pupils from different schools commented that they preferred learning online as they think this is the future of learning and how they will communicate when they are older. When asked why one pupil said:

"When I am at home, I am always using the internet for my homework and I often Skype my older brother to help with my homework."

When asked if they could explain about the future of learning being online, the pupil said:

"All my friends have mobile phones and gaming platforms like PlayStation and Xbox's, most nights I speak with them online playing games. I think as I get older we get better computers to make learning easier."

Almost all pupils asked in the focus groups have used the internet in the last month to communicate with friends or family. Gaming was frequent response to how pupils speak with friends online, with the majority owning a headset device to communicate. Yet, a few pupils mentioned that they would prefer to be taught in person rather than online. When asked to explain why, a pupil responded:

"I would prefer to have the teacher in the room, as it is easier if they can show me by looking at my work and they can mark my work."

While discussing how the lessons were designed on the platform, a number of pupils commented that they would often wait until another pupil had typed in their answer before they wrote on the screen. A pupil commented:

"I just waited for X to type her answer as she always gets the answer right."

When asked if they understood how to the answer the question they replied:

"No, but I see the answer of the other pupils, so I just type this in."

Lesson observations also identified a number of learners waiting for their peers to answer a question before typing their own, often without working out the mathematics problem on paper. The issue was more prevalent with a particular teacher, with the other teacher using strategies such as counting down 5, 4, 3, 2, 1 then asking all pupils to answer their questions at the same time.

On the whole, the majority of pupils engaged in the intervention and had a positive learning experience. The focus of the feasibility study involved the methodological development of using aggregated small scale RCTs to link to a cumulative prospective meta-analysis. In future feasibility studies, using questionnaires as a measure of engagement and to capture high quality feedback from pupils on eth intervention should be incorporated into the design of the process evaluation.

<u>Teacher views</u>

An emergent theme from the interviews with teachers was that of overcoming the logistical barriers for implementing the intervention. Teachers were generally positive in the engagement of learners with the online small group teaching, yet the logistical barriers were a potential concern if the intervention was to be deployed as a long-term solution. The issue was highlighted by teacher C, saying:

"I was really happy with the pupil engagement in the intervention, as I was expecting to have to chase the learners to attend the lessons. However, finding a free computer room during the school day was difficult due to timetabling constraints."

When prompted to explain further on the timetabling constraints, the teacher responded:

"The school has reduced the number of computer rooms and purchased tablets for departments. As the software needs to be on PC's, I really struggled to find a room that was available for the online mentoring."

Teacher B provided further insights into the logistical issues:

"When the intervention worked well, it was great. However, finding a computer room with PCs that worked was an issue as most of the main computer rooms are scheduled during lesson times."

"Also, as the intervention still requires a person to supervise the learners in a computer room I have had to remove the teaching assistant support from a class."

The comment regarding removing additional support is important, as a spill over effect could occur and impact learners not participating in the programme. In addition, if learners are removed from a maths lessons a smaller class teacher to pupil ratio could provide a positive learning spill over effect.

A second emerging theme from the interviews was behaviour of pupils during the online lessons. On three schools, the off-time task noted by the observer was minimal with pupils engaged in the intervention. All teachers highlighted the positive engagement, with teacher A saying:

"It is clear the students engaged as the usual off task chatter was minimal in all the lessons. As soon as the headsets were on, the room went silent and the pupils were listening to the teacher."

Further comments linked to the positive behaviour include:

> *"Usually with extra maths interventions, it can be a battle to get the pupils to attend. Over the 5 weeks, I only had to chase a small number of pupils who had genuinely forgot about the lessons. This was a major positive for the intervention in our school."*

> *"Student A (name removed for anonymity) is usually late for maths, but he was always one of the first to be collecting the headsets at the start of the lesson".*

When asked about Student A and if the teacher had seen a difference between the online and traditional classroom setting, they commented:

> *"Yes, usually he comes into the lesson and is off task talking with friends within the first few minutes. As soon as the headset is on, he goes into a zone and does not get distracted so easily."*

The lesson observations and TUTE teacher feedback supports the view that students positively engaged in the intervention with the exception of lesson observations in school 3. It is important to acknowledge that the teacher interviews or focus groups did not involve school 3, due to time restrictions and availability of the school to complete focus group and interviews. A further consideration is the presence of an observer in 18 of the lessons, as pupils may be inclined to change their behaviour when being observed.

7.5.5 What are the staff perceptions of the current and possibly sustained impact of the intervention on children's educational attainment?

The teacher's perceptions of the current and possibly sustained impact of the intervention were mostly positive. An emergent theme related to the perceived impact involved the engagement if the

learners in the intervention. Teachers appeared to link pupils sitting quietly with the potential impact of the intervention.

> *"The pupils were listening to the lessons and most of the pupils were able to explain to me what they had learned after when I questioned them."*
>
> *"The students were well behaved and appeared to learn from the lessons, I will be interested to see the post test data to see the impact of the intervention."*
>
> *"I think the intervention has had a positive impact on the pupils, I was impressed with their behaviour and from the lack of noise I think the intervention was a success."*

The perceived learning and engagement factor is similar to a recent trial conducted for increasing literacy in primary pupils. Chatterbooks is an extracurricular reading initiative that aims to increase a child's motivation to read by providing schools with tools and resources to encourage reading for pleasure. Teachers had recommended the intervention for an independent evaluation as pupils enjoyed the programme. However, an independent trial found a negative effective size of -2 months compared to the control (Styles, Clarkson & Fowler, 2014).

7.5.6 What are their suggestions for change if the intervention was to be more widely implemented?

The emerging themes from the interviews with teachers were around mobile delivery (tablets) and finances. Firstly, all three teachers identified a move towards their schools purchasing iPad or tablet style devices for departments, often at the expense of converting department computer rooms to classrooms.

> *"Over the last few years, our school has purchased tablet devices for our department as the computer room has been converted into a traditional style classroom. We store the iPad in a*

case and this is booked by the teachers in the department when we require access to devices".

When asked if it would be easier for lessons to be delivered on these devices, the teacher responded:

"Definitely, as I could use any free available space in the school for the intervention and it would be easier the pupils to write on the tablet screen rather than use the mouse to draw."

The teacher elaborated on the mouse feature when prompted, saying:

"The online classroom is great but with mathematics it is often easier to write out the solution rather than type."

Furthermore, the teacher explained:

"If the intervention was to be more widely implemented from homes, as it requires a PC / laptop the business might find that this limits the uptake… due to learners sometimes only having access to mobile style devices in their home."

The second theme relating to finances is also fundamentally important for wider implementation of the intervention in schools. Government funding has been on a downward trend for schools over recent years, as demonstrated by a recent report by the Institute of Fiscal Studies.

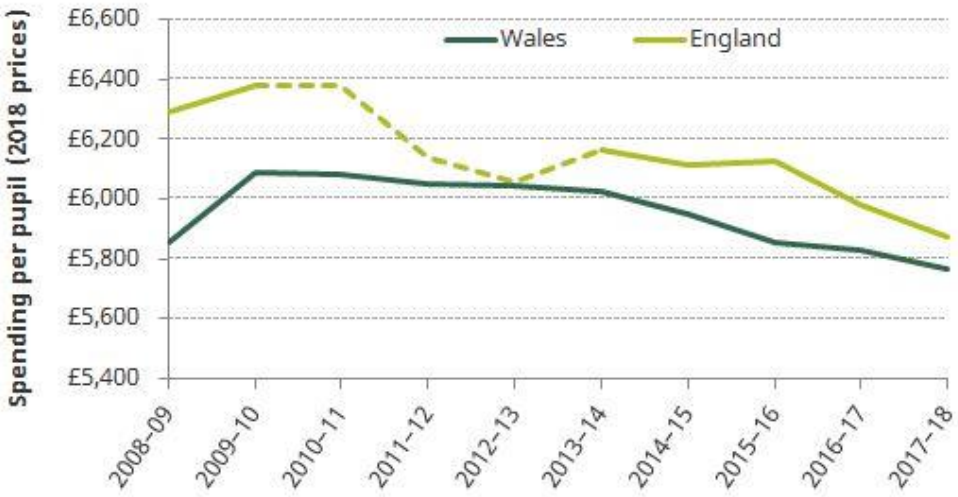


Figure 26: School spending per pupil in England and Wales between 2008–09 and 2017–18 (Sibieta, 2018)

All three teachers referenced the cost of the intervention as a potential issue for wider implementation within schools.

> *"The intervention is great, but my only concern is that our budget is used for exercise books and a few other resources. We do not have the funding within the department to fund such an intervention at present."*

The cost per lesson per pupil for the intervention is currently £5, however with 12 pupils in a lesson the cost per hour is £60 excluding licence agreements. Unless schools use additional funding, such as the pupil premium, scaling the intervention to become more widely implemented by be difficult.

7.6 Conclusions

The headline finding is the TUTE online mathematics small group teaching is a successful approach in this context of this study.

7.6.1 Limitations

The initial design of the study involved three cohorts of schools completing data collection over an academic year. Unfortunately, only one cohort of school data was completed due to platform provider deciding to focus on a qualitative engagement study rather a randomised controlled trial design. Consequently, the study is limited in the generalisations that can be made from the small sample size.

One school selected pupils from a lower ability group compared to the remaining four schools, also reducing the intervention time from 45 minutes to 30 minutes. It is highly likely that this impacted the fidelity of the study and had a negative impact on the overall effect size of the intervention.

The design adopted for this trial cannot consider any long-term impact of this intervention. Once the intervention was completed, the control group in the schools were offered the opportunity to receive the intervention.

The schools participating in the trial had volunteered to participate in the trial as they had already participated in the online cross-age peer mentoring study, except for a school who was already receiving online lessons from TUTE The schools are therefore not necessarily a representative sample of a larger population.

TUTE developers were not involved in this evaluation of the intervention which means that there was no conflict of interest.

The results of AlfieSoft assessment are based on pupils' performance on the screen-test. AlfieSoft is a computerised on screen-test and is independent of the TUTE lessons. Therefore, the self-marking software reduces the possibility of bias between the intervention and control groups.

7.6.2 Key feasibility conclusions

The study demonstrated the ability of a single researcher to conduct a small scale randomised controlled trial with the intention to replicate the study across a further two cohorts to allow the aggregation of the data for a prospective cumulative meta-analysis.

7.6.3 Impact evaluation conclusions

The main objective of this RCT was to test the gains in mathematics ability against those of pupils in the control group. The criteria of a trial were maintained throughout and no schools dropped out of the evaluation, the TUTE intervention was delivered over a 5-week period, the pupils in the control did not have access to the TUTE system and the attrition rate was 5%.

The trial was small in scale as only one cohort of data was able to be collected, therefore the effect size must be treated with caution. The overall finding suggests that TUTE was an effective mathematics intervention with the effect size of around +0.44. However, at school level the effect size ranged from -0.57 to +0.74, demonstrating the need for caution with such as small sample size.

7.6.4 Process evaluation conclusions

The process evaluation confirmed that the intervention was able to be delivered with a high degree of fidelity, however it is important to note that the high degree of researcher involvement (18/24 sessions observed) provided additional support that would not normally be available to schools implementing the intervention. Pupil and teacher views were mostly positive towards the perceived impact of the intervention and attendance in the lessons was good. Three main barriers to implementation and wider were identified; technical, logistical and finances. The findings would inform the next two stages of the aggregated trial if these had continued to be delivered as originally planned.

7.6.5 Summary

Table 29: Overview of Study 2 research questions, methods, data and analysis.

Research question	Method	Data	Analysis	Conclusion
What is the effectiveness of small group online mathematics programmes compared with "business as usual" on the maths skills of participating children?	Randomised Controlled Trial	Pre-test, post-test ALFIE-soft data	Multi-level model (linear mixed effects fit by REML)	Positive effect size 0.44 (95% CI – 0.3to 1.17). Unable to generalise due to sample size.
Does the dosage analysis impact on the effect of the intervention?	Randomised Controlled Trial	Pre-test, post-test, dosage of intervention.	CACE analysis	Programme only included 5 lessons, limited evidence.
Is there a differential impact on pupil premium students?	Randomised Controlled Trial	Pre-test, post-test.	Multi-level model (linear mixed effects fit by REML)	Unable to complete due to small sample size
Is it feasible to implement a prospective cumulative meta-analysis for the aggregation of small scale randomised controlled trials in schools?	Process evaluation (observation, interviews and focus groups)	Lesson observation, TUTE teacher feedback and fidelity score	Comparison of the fidelity scores and lesson observation analysis	Unable to complete due to only one cohort of schools

Chapter 8: Development of trial protocols

This chapter outlines the reasoning for the use of protocols for the primary empirical studies in education. The next section looks at the use of protocols in research, focusing on clinical trials and education. The third section presents a framework for the development of protocols for a prospective cumulative meta-analysis (PCM). The framework for the development of protocols is followed by an explanation and elaboration section. This provides important information to promote the full understanding of the checklist recommendations with a rationale, description and references supporting its importance. The chapter concludes with a discussion on the need for high quality protocols in educational research.

8.1 Why use protocols in research?

This section provides a working definition for the term protocol, a review of the use of protocols in clinical trials, a review of protocols in educational research and a discussion of the lessons which can be learned from the literature.

8.1.1 Defining the term protocol

The definition of a protocol, according to the English Oxford Dictionary is "the accepted or established code of procedure or behavior in any group, organization, or situation" (University Press, 2016). When applied to research in clinical trials, Chan et al. (2013) acknowledge that the precise definition of a protocol varies amongst different stakeholders. However, in the process of creating the SPIRIT 2013 guidelines, the authors defined a protocol as "a document that provides sufficient detail to enable understanding of the background, rationale, objectives, study population, intervention, methods, statistical analyses, ethical considerations, dissemination plans, and administration of the trial; replication of key aspects of trial methods and conduct; and appraisal of the trial's scientific and ethical rigour from ethical approval to dissemination of results" (Chan et al., 2013).

For the purpose of this PhD, the term protocol will be defined using the SPIRIT 2013 definition with the added statement of providing a living document for researchers to use, ensuring transparency and rigour for educational research. It is important that any changes from the protocol are described in publications or reports, with justifications for any amendments clear for the reader to understand the reasoning for any deviations from the original plan. In research, such deviations should be expected due to the complex nature of the settings of the research and participants involved. It is

important to consider at which point the changes to the original protocol result in the study to be considered a different project from the initial research. In the primary research for study one for online peer tutoring, if the PCM established sufficiency and new research questions are asked such as does online peering group size effects replicate with a different area of mathematics, such as algebra instead of the initial data handling topic, then a new project is initiated.

As the definition of a protocol implies, the purpose of the document in clinical trials is to facilitate an appropriate assessment of the ethical, safety and scientific issues before a trial begins, implement consistency and rigour during trials and allow an assessment of consistency between the original intent and final report (Dwan et al., 2008). The proceeding part of this chapter reviews the development and use of protocols in clinical trials.

8.1.2 A review of the use of protocols in clinical trials

The use of RCTs as a method of assessing efficacy of health care interventions has been established for a number of years. However, the guidance on the reporting of clinical trials had been limited with a growing number of studies concerned about the methodological quality in the reporting of RCTs (Altman & Dore, 1990; Mahon & Daniel, 1964; Pocock, Hughes, Lee, R.J, 1987). In 1993, a workshop was held in Ottawa, Canada, with the purpose of developing a scale to assess the methodological quality of RCTs in healthcare. Although the aim of the initial meeting was to develop a scale to assess the methodological quality of RCTs, preliminary discussions between the 30 academics suggested that the initial purpose of creating the scale was irrelevant due to the lack of reporting of these methodological items in published reports. A unanimous decision was reached to focus the workshop on guidelines to improve the reporting of RCTs in healthcare interventions, resulting in the publication of Standardised Reporting of Trials (SORT) statement. The SORT statement set out 24 essential items that were required in the reporting of a trial, with supporting empirical evidence to explain why these items should be included and a recommended format on how to include these (Andrew, 1994).

In addition to the SORT statement, in 1993 and independent of the meeting in Ottawa, Canada, another group were meeting to discuss similar issues facing the reporting of clinical trials in the biomedical sciences. This group, the Asilomar Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature, developed a similar checklist of items to include when reporting clinical trials (Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature, 1994). The stimulus for the creation of these two independent workshops had been the growing evidence and concern regarding the reporting of RCTs in healthcare over a number of years. In a review conducted by Freiman et al. (1978) into the sample sizes involved in 71 “negative”

RCTs, they found many therapies labelled as “no different from control” in trials which used sample sizes that were inadequate and had not received a fair test. As a result, the authors called for more attention in the planning of clinical trials for sample sizes, as a concern for missing important therapeutic improvements were a possibility due to inadequate planning of these trials. Further research by Moher et al. (1994) supported the conclusions provided by Freiman et al. (1978) with most trials with negative results not having a large enough sample to detect a 25% or 50% relative difference. The authors concluded that the reporting of statistical power and sample sizes needed to be improved in the planning and reporting of RCTs in healthcare.

In 1995, representatives of both groups met in Chicago, USA with the aim of creating a single document based on the SORT and Asilomar proposals for guidance on the reporting of clinical trials. As a result of this meeting, the Consolidated Standards of Reporting Trials (CONSORT) Statement was published in 1996 (Begg et al., 1996). The CONSORT statement consisted of a 21-item checklist pertaining mainly to the methods, results and discussion of an RCT report, allowing the publication of the important information needed to evaluate the internal and external validity of a trial report. Additional meetings of the CONSORT group in 1999 and 2000 resulted in a revised CONSORT statement in 2001 (Moher, Schuluz, Altman, 2001). The revised statement was supported by the development of an explanatory document to enhance the use and dissemination of the updated guidelines. In 2001, and without precedent, the revised CONSORT statement was published simultaneously in three prestigious international medical journals, the Lancet, JAMA and Annals of Internal Medicine.

As a result of an increasing evidence base, the CONSORT group regularly updated the statement with further iterations of the guidelines in 2007 and 2010. Empirical studies highlighting new concerns, such as outcome reporting bias (Chan et al., 2004) are addressed in the recent versions of the statement. Furthermore, the release of the 2010 CONSORT revised statement (Schulz, Altman, Moher, & Group, 2010) was simultaneously published in eight leading journals, demonstrating the significance of the CONSORT guidelines for trial reporting for healthcare interventions.

The introduction of the CONSORT guidelines in healthcare has addressed many of the concerns raised for the transparency and reporting of methodological items in RCTs, yet this has only increased the clarity of the reporting and not the actual conduct of a study. As described by Fiona Godlee and Trish Groves from the British Medical Journal, “the reporting guidelines are the ambulance at the bottom of a cliff, rather than the fence at the top” (Schulz & Grimes, 2013). The reference to the “fence at the top” alludes to the primary purpose of a quality protocol. Unless a protocol includes a transparent and clear outline of the intended research, it is difficult for external reviewers to assess the consistency between the intended research and published report. The significance of this is

supported by a growing evidence in clinical trial literature for the variation in the quality of trial protocols (Chan et al., 2004; Dwan et al., 2008; Pildal et al., 2005; Smyth et al., 2011; Tetzlaff et al., 2012). The following section will now discuss the importance of the evidence of protocol deficiencies in clinical trials, exploring the consequences for research and the development of SPIRIT 2013.

8.1.3 A review of the evidence for protocol deficiencies in clinical trials

The purpose of a trial protocol is clear; to describe the rationale, methods, analysis plan and administrative details from the inception of the trial to the dissemination of the findings. The importance of clear and transparent protocols are fundamental requirements for all trials, as various stakeholders from researchers, ethics committees, funding bodies, journal editors and external reviewers rely on the validity of the information reported. The consequences of poor or lack of reporting can result in bias in the final trial results, with resources and time potentially wasted on interventions based on incorrect data. The ethical consequences of decisions made based on biased findings could also impact the health of patients, as ineffective treatment programmes could be continued or introduced with negative implications for patients.

A primary concern raised in the CONSORT working groups has been the issue of outcome selection bias (ORB) in reporting RCTs. Within-study selective reporting bias involves studies that have been published and Dwan et al. (2013) define this as "the selection on the basis of the results of a subset of the original variables recorded for inclusion in a publication". For example, if a study protocol involved 5 primary and 10 secondary outcomes, the authors find significant results for only one primary outcome and three secondary outcomes, the final publication focuses on the significant findings. The lack of reporting of the non-significant outcomes in the final publication would bias the results and without clear and complete protocols, it would be difficult for external reviewers to discover these deviations from the initial research aims. Consequently, with RCTs forming the primary evidence for systematic reviews and meta-analyses, lack or incomplete reporting can affect the validity of future findings. An important issue to consider is the impact of representation bias in meta-analyses. The bias that arises from non-published outcome data missing outcome data that can affect a meta-analysis can be on two levels: study level problem whereby non-publication by researchers or rejection from publishers and an outcome level problem. The latter involves the selective non-reporting of outcomes within the published results, with deviations from the primary protocol introducing a bias that is likely to overestimate the effect of the experimental treatment.

Chan et al. (2004) conducted a cohort study using protocols and published reports of RCTs approved by the Scientific-Ethical Committee for Copenhagen and Frederiksberg between 1994 and 1995. The primary objective of the study investigated the extent and nature of outcome reporting bias

in a cohort of randomised controlled trials. The study sample consisted of 102 trials with 122 published journal articles, with 3736 outcomes identified. In respect to the prevalence of outcome reporting bias, the authors found 62% of trials had at least one primary outcome that was changed, introduced or omitted when comparing published articles with protocols. In response to the survey regarding these discrepancies, 86% of trial co-ordinators (42/49) denied the existence of unreported outcomes despite the clear evidence for this. Furthermore, statistically significant outcomes had higher odds of being fully reported compared to non-significant outcomes. It is important to acknowledge the potential limitations of the study, as the RCTs involved were initiated in the years 1994 to 1995, before CONSORT guidelines were fully accepted as best practice in reporting RCTs. Also, the sample of 102 RCTs is limited to Denmark, with no clear explanation as to why this region was selected as the sample. Yet, the findings suggest that in this cohort of studies the reporting of trial outcomes was not only incomplete but biased and inconsistent with the intended outcomes in their protocols. Chan et al. (2004) conclude that the published articles and meta-analytic reviews that incorporate them may therefore be unreliable due to an overestimation of the effectiveness of the intervention.

Pildal et al. (2005) conducted a review using the same sample of 102 trials from the protocols of RCTs approved by the scientific and ethical committee for Copenhagen and Frederiksberg used previously by Chan et al. (2004). The main outcome involved the frequency of adequate, unclear and inadequate allocation concealment and sequence generation in trial publications compared to protocols. A second primary outcome involved investigating the proportion of protocols where methods were reported to be adequate but descriptions were unclear in the trial publications. Ninety six of the 102 trials had unclear allocation concealment according to the trial publication, with 81 out of these 96 trials having unclear or inadequate concealment in their protocols. Interestingly, a study by Mhaskar et al. (2012) assessing whether reported methodological quality of RCTs reflects the actual methodological quality outlined in protocols found higher methodological quality in protocols compared to published reports. The sample included four hundred twenty-nine phase III RCTs published by eight National Cancer Institutes between 1968 and 2006. Once again, potential implications of these findings for meta-analyses are that the assessment of the quality of the effect size based on reported quality alone can produce misleading results.

In 2010, a review of the frequency and reasons for outcome reporting bias in clinical trials was conducted by Smyth et al. (2011). The study compared trial protocols with subsequent publication(s) to identify any discrepancies in the reported outcomes. The study involved 268 trials, with 183 identified from the Cochrane systematic reviews where at least one trial was suspected of being at risk from reporting bias and 85 from a random sample of PubMed journal between August 2007 and July 2008. Telephone interviews were conducted with the respective researchers to develop an

understanding into the issue of selective outcome reporting. From the initial sample of 268 trials, 161 researchers responded to the request but a failure to arrange contact or obtain a copy of the protocol resulted in only 59 interviews being conducted (37%). The study found that the prevalence of incomplete outcome reporting was high across the sample, with researchers unaware of the possible implications for not reporting protocol changes and all outcomes in the published studies.

A review by Blümle et al. (2011) investigated whether and how eligibility criteria of participants pre-specified in protocol of RCTs are reported in published articles. The sample included protocols submitted to the research ethics committee of the University of Freiburg in the year 2000. The data source included 52 trial protocols and 78 subsequent publications published between the years 2000 to 2006. The main outcome of the study was to determine if the eligibility criteria in the published article were either matching, missing, modified or newly added from the original protocol. The review found that discrepancies were found between the published articles and protocols for all 52 studies. The authors concluded that the published articles in the sample generally did not reflect the pre-specified study population in the protocol.

Finally, a systematic review of the empirical evidence of study publication bias and outcome reporting bias by Dwan et al. (2013) reviewed the evidence available on this particular issue. The review included twenty cohort studies, fifteen investigating publication bias and five focused on outcome reporting bias. With regards to the outcome reporting bias, the authors found that 40 – 62% of studies had at least one primary outcome that was changed, introduced or omitted when comparing publications to protocols. A meta-analysis was not possible due to the differences between the studies.

In sum, the evidence confirms the earlier conclusions from Chan et al. (2004) on the prevalence of outcome reporting bias in the reporting of RCTs for healthcare interventions. The introduction of the CONSORT guidelines can be linked back to the analogy by Godlee and Trish Groves as the ambulance at the bottom of the cliff, yet the fence at the top was an issue that needed to be addressed in order to provide confidence in the validity of research findings.

8.1.4 The creation of SPIRIT 2013 protocol guidelines

The Lancet published an editorial feature titled “Strengthening the credibility of clinical research” (The Lancet, 2010) in response to a two year investigation by the United States Senate Committee into the excessive cardiovascular events in patients taking a drug manufactured by GlaxoSmithKline (GSK) called Rosiglitazone. The inclusion of the RECORD journal article published by Lancet in 2009 was at the centre of storm, with allegations of intimidation of researchers and manipulation of trials for financial gain. The Senate found evidence of misconduct in the unblinding of the trial two weeks before suggesting an interim analysis, delaying the withdrawal of the drug for financial reasons. The Lancet conducted an internal audit, comparing protocols with subsequent Lancet publications. As the evidence outlined previously in the chapter has demonstrated, the Lancet found that adherence to pre-published protocols was selective and ambiguous (The Lancet, 2010). In a defence of the credibility of clinical trial literature, the editorial team at the Lancet outlined the case for better designed, freely accessible and more transparent protocols. The article illustrated the important role CONSORT guidelines had transformed the reporting and design of clinical trials, before introducing the SPIRIT statement as a solution to the issue of improving the design and robustness of trial protocols.

In 2007, SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) initiative was launched with the primary aim of improving the content of trial protocols with the publication of the SPIRIT 2013 as a guideline for the minimum content of a clinical trial protocol (Chan et al., 2013). The SPIRIT 2013 statement was developed through a consultation of 115 key stakeholders involved in clinical trials. The diverse nature of the stakeholders involved in clinical trials as outlined in the development of the SPIRIT 2013 statement, highlights the need for transparent and clearly written protocols. After numerous iterations, the 33-item checklist provided guidance to facilitate the creation of high-quality protocols in clinical trials, with the aim of increasing the completeness and transparency for the benefit of all stakeholders. Appendix X provides an example of the SPIRIT 2013 Checklist: Recommended items to address in a clinical trial protocol and related documents.

In addition to the SPIRIT 2013 guidelines, the authors published the SPIRIT 2013 explanation and elaboration paper to provide a rationale for each item on the checklist with empirical evidence and examples of best practice (Chan et al., 2013). The development of the guidance for constructing and reporting protocols should see a greater transparency and clarity when assessing the quality of research in clinical trials.

In conclusion, significant evidence exists in clinical trial literature regarding the reporting of RCTs in healthcare interventions and the discrepancies between published literature and the initial

intended research outlined in study protocols. The development of CONSORT has increased the quality of trial reporting, yet issues such as outcome reporting bias can still call into question the validity of research findings. The introduction of SPIRIT 2013 is a move towards a more transparent process of trial reporting, strengthening the credibility of research again for healthcare interventions.

The next section of this chapter reviews the current status of the use of protocols involved in educational interventions using RCTs, what guidance exists and if the lessons have been learned from the experiences of clinical trials in healthcare with a discussion on the implications for the validity of educational research involving RCTs, systematic reviews and meta-analyses.

8.2 A review of the use of protocols in educational research

As education moves increasingly into an evidence-based practice, the need to be able to assess the methodological quality of RCTs for informed education decision making is crucial if policy makers, school leaders and teachers use this evidence to inform practice. Therefore, consumers of research evidence need to make decisions on the basis of their confidence in the methodological quality of research published and without the publication of protocols in education it is impossible to assess the consistency between the original intent of the research and the final publication or report.

In educational research, few clear guidelines on the reporting of trial protocols are evident in the research literature. It is not possible to ascertain the frequency and quality of trial protocols in educational research without a systematic review of the evidence. However, a common theme occurs when scoping the literature for protocols in education. In the search of RCTs acknowledging the publication of protocols prior to the start of data collection in education, trials funded through the EEF or other funding bodies and published in academic journals provided links to the protocols on the funders website (Gorard, Siddiqui, & See, 2016; Torgerson et al., 2016; Siddiqui, Gorard, & See, 2015). Independent trials funded by the EEF are required to publish evaluation protocols on their public website, using a guide with items based on the 2010 CONSORT statement (EEF, 2016). The protocol supports evaluators in considering the key features such as a description of the intervention, significance, research questions, design, randomisation, participants, outcome measures, sample size calculations, analysis plan, implementation and process evaluation methods, costs, ethics and trial registration. The requirement of a published protocol by the EEF ensures a transparency between the original intent and the completed research report.

The publication of protocols in research journals relating to education is not common, although recently a few protocols have been published (Have et al., 2016; Allen Thurston et al., 2015). Unfortunately, many published journal articles using RCTs in educational research do not include or

require the publication of a protocol. The absence or the publication of incomplete protocols has a negative impact on the process of external review. As RCTs regularly form the basis for systematic reviews and meta-analyses, the lack of and incomplete reporting of the methodological decision making creates a difficulty for any researcher attempting to synthesise the data. As a result of the evidence from clinical trials literature (Blümle et al., 2011; Chan et al., 2004; Dwan et al., 2008; Pildal et al., 2005; Tetzlaff et al., 2012b), it is evident that even use of protocols does not ensure transparency and high quality guidance is required to fulfil the intended purpose of these protocols. Therefore, if RCTs are developed and implemented in education without the publication or creation of a well written protocol, how is it possible for external reviewers or stakeholders to able to appraise the conduct or reporting of these trials? If journal publications do not require the publication or submission of protocols, the question must be raised regarding the validity of published articles and to what extent outcome reporting bias and other related biases linked to deviations from intended research are present. Currently, educational research has few clear guidelines for trial reporting or the construction of quality protocols, access to protocols or the trial registration of all educational RCTs, educational research finds itself in a muddied mess. As the focus of policy highlights RCTs as the gold standard for inferring effectiveness for educational interventions we find ourselves stuck in a loop of creating evidence whereby we have no safeguarding or fence at the top of the cliff. As the Lancet editorial feature explained, "while nothing can completely protect against scientific misconduct, SPIRIT – in addition to improving overall trial quality – will make deviations more difficult to execute and disguise" (The Lancet, 2010).

8.2.1 What lessons can be learned from the development of protocols in clinical trials

In educational research, it is evident that a distinct lack of specific reporting guidelines exists for the reporting of RCTs and in the development of high-quality protocols. In a review comparing the methodological quality of RCTs in health and education, Torgerson et al. (2005) undertook a methodological comparison between healthcare and education trials involving 96 placebo drug trials, 54 non-drug trials, 54 trials in specialist health journals and 84 trials in educational journals. The authors reported that while the reporting of healthcare trials improved over time, the reporting quality of RCTs in education declined. This lack of quality in the reporting of methodological quality in educational trials creates issues for those who use this evidence in systematic reviews and meta-analyses. Torgerson et al. (2005) recommends the development of a checklist similar to the CONSORT statement but specifically for educational trials, enabling educational journal editors a framework to quality assure the publication of educational trials. Yet, it should be argued for the creation of guidelines for protocols if researchers are using randomised controlled trials in education. If ethical

boards in Universities required the creation and submission of well-defined protocols, then editors at journals would be able to request protocols during the peer review process, publishing these as supplementary material upon publication of the research. The use and publication of protocols by peer reviewed journals would allow external reviewers greater transparency and reduce multiple sources of bias.

At present, the lack of a requirement for the development of trial protocols and analysis plans means that researchers have little incentive, apart from professionalism, in submitting these. The pressure for peer review in educational research creates a system whereby the risk of potential bias can thrive, reducing the robustness of evidence. There are compelling ethical reasons for the inclusion and publication of trial protocols in education, so the question must be raised as to why these are not included at the conception of any educational trial. A plausible explanation for the lack of the creation of protocols and analysis plans may be a lack of understanding about why these are needed, as many might know about RCTs but they do not have sufficient experience to know why a protocol is needed. Therefore it might not be the lack of incentive but rather a lack of experience that results in the lack of submissions.

8.3 Framework for the development of protocols for a prospective cumulative meta-analysis

As part of this thesis, following framework has been created as a model of good best practice for the creation of trial protocols in the initial stages of the design of aggregated trials linked to a prospective cumulative meta-analysis. The framework has been adapted from the SPIRIT 2013 (Chan et al., 2013), CONSORT statement (Schulz, Altman, & Moher, 2010) and the guidance from the EEF (EEF, 2016). The purpose of the checklist 2016 is to ensure each individual trial cohort included in the prospective cumulative meta-analysis is rigorous and transparent from the conception of each trial.

Table 30: Checklist 2016: recommended items for educational randomised controlled trials protocols in education

Section	Item	Description
Administration		
Title	1.1	Descriptive title
Trial registration	1.2	Trial registry name and identifier
Protocol version	1.3	Version identifier and date
Funders	1.4	Identification of funders, resources and additional support
Roles and responsibilities	1.5	a) Names, institution and roles of protocol contributors
		b) Contact details of trial funders
		c) Outline of the roles and responsibilities of all personnel involved in the trial
Significance of study		
Intervention	2.1	Detailed description of the intervention including TIDieR items (if possible)
Significance	2.2	Background and rationale of the research study
Theory of action or logic model	2.3	Theoretical frameworks underpinning the research
Research questions	2.4	Clear research questions to be answered by the trial, defining its purpose and scope.
Methods		
Trial design	3.1	Description of the trial design
Participants	3.2	a) Inclusion and exclusion criteria for participants
		b) Description of study settings and locations where data will be collected. Reference to where a list of study sites can be located.
Randomisation	3.3	a) Description of method used to generate the random sequence
		b) Type of randomisation, including details of any restrictions (such as blocking and block size)
		c) Description of allocation concealment mechanism (e.g. sequentially numbered opaque envelopes / containers) and steps taken to conceal the sequence until interventions are assigned
		d) Implementation details on who will generate random allocation sequence, who enrols participants and who assigns participants to the intervention

		e) If completed, who will be blinded after assignment to interventions (for example, participants, those assessing outcomes) and how
Outcomes measures	3.4	Defined pre-specified primary and secondary outcome measures, including specific measured variable (raw score, age standardised score or gain scores)
Sample size	3.5	Description of estimated number of participants needed to achieve study aims and how this was calculated, including statistical assumptions supporting sample size calculations
Recruitment	3.6	Outline of strategies to be used to achieve the adequate sample size
Data management		
Data collection methods	4.1	a) Plans for assessment and collection of baseline, outcome and other trial data.
		b) Description of study instruments along with reliability and validity (for example Alpha Cronbach / RASCH analysis)
Participant flow diagram	4.2	Clear diagram with number of participants who were randomly assigned, received intended treatment and analysed for the primary outcome
Data retention	4.3	Plans to promote retention of participants, including plans for data collection for participants discontinuing from the protocol
Statistical methods	4.4	a) Description of statistical methods for analysing primary and secondary outcomes
		b) Methods for any additional analyses (for example, sub group or adjusted analyses).
		c) Description of any statistical methods that will be used to handle missing data
Implementation and process evaluations		
Research questions	5.1	Summary of research questions to be addressed by the evaluation and process evaluation (if possible, link back to the logic model or theory of action
Methods	5.2	Description of methods used to address the research questions for the implementation and process evaluation
Fidelity	5.3	Outline on how the programme fidelity will be assessed
Ethics and dissemination		

Ethics approval	6.1	Name and details of the ethics committee responsible for storing the protocol (if a suitable trial registry is not available)
Protocol amendments	6.2	Plans for communicating important protocol modifications to relevant stakeholders
Confidentiality	6.3	Description on how personal information from participants will be collected, stored and maintained in order to meet data protection standards
Dissemination	6.4	Plans for researchers and sponsors to communicate trial results to relevant stakeholders, including plans for access to the full protocol.
Appendices		
Ethics materials	7.1	Model consent forms, school agreement to participate agreements and other related documents
Other related documents	7.2	Assessment materials (if available with copyright permission)

The checklist 2016 includes 28 items that should be considered when designing a protocol for the small-scale aggregated trials in education. A brief explanation and rationale for each item will now be explained. Many items are self-explanatory and are accepted as best practice in conducting RCTs, however these are often missing in trial protocols.

Administration

1.1 Title – A descriptive title is important as this should be a succinct description of the topic and basic study design. This is important in trial identification and retrieval in literature searches, systematic reviews and meta-analyses.

1.2 Trial registration – There are compelling scientific and ethical reasons for trial registration, yet no requirements are enforced for trial registration in educational or social science peer reviewed publications. The registration of all RCTs at conception would increase transparency and ensure all trials are included in relevant literature searches, decreasing the risk of publication bias in systematic reviews and meta-analyses.

1.3 Protocol version – Publication of the trial protocol, sequentially labelled and dated with a protocol version number. It is important for each version to include an explicit listing of changes made to the previous protocol, allowing external reviewers a transparency in the decision making throughout the trial.

1.4 The identification of funders, resources and additional support – Ginsburg & Smith (2016) identified developer association with RCTs as a potential threat to the validity, as a review of 27

RCTs for the WWC found that 44% of the author reports had an association with the curriculums developers. Therefore, a protocol should be explicit in the identification of trial funders, resources and additional support to allow a greater transparency.

1.5 Roles and responsibilities – This information helps to ensure a clear understanding of the roles and responsibilities at the onset of the trial.

Significance of the study

2.1 Intervention – A detailed description of the intervention should be included to prompt authors to describe the intervention in sufficient detail to ensure replication is possible, as this is a fundamental requirement of the protocol in the design of aggregated trials linked to a PCM. Authors should include a Template for Intervention Description and Replication (TIDierR) checklist to describe the intervention. Hoffmann et al., (2014) highlight the importance of using TIDieR to improve the reporting of interventions, ensuring greater transparency for external reviewers and researchers looking to replicate the study.

2.2 Significance – The significance of the study is important for scientific and ethical reasons for undertaking the trial, with a summary of the literature (published and unpublished) undertaken. It is recommended that the author places the trial in the context of the available evidence, ideally with an up-to-date systematic review of relevant studies.

2.3 Theory of action or logic model – Theoretical frameworks underpinning the research should be discussed as either the theory of action or logic model. Effective logic models make an explicit, often using visualise representations, statement about the processes that will bring about change and the results you expect.

2.4 Research questions – Clear research questions are outlined and relevant to the design of the trial, defining its purpose and scope.

Methods

3.1 Trial design – A clear description of the design including the type of trial used. It is important to explain the choice of study design and why this is the best design to answer the research questions.

3.2 Participants – Inclusion and exclusion criteria for the participants should be clearly outlined, as this is a strength of the use of protocols for aggregated trials linked to a prospective cumulative meta-analysis. The nature of the PCM ensures the inclusion criteria reduces the variability

between the trials, as this often a concern in traditional meta-analyses (Cochrane, 2016). Additionally, a description of the study settings and locations of data collected should be referenced.

3.3 Randomisation – The methods used to generate the random allocation should be explained, with reasoning for pairing, stratification or minimization. Additionally, the protocol should include the plans for recording the randomisation process.

3.4 Outcome measures – The outcome measures should be pre-specified for the primary and secondary outcome measures. Details of any plans to ensure the tests are blinded should be included in the protocol.

3.5 Sample size – The protocol should include how the sample size is determined with power calculations, explaining the assumptions and restrictions. For example, if the capacity of the developer is a limiting factor this must be noted. Among randomised trial protocols in clinical trials that describe a sample size calculation, Chan et al. (2008) found between 4 – 40% of studies do not state all the components of this calculation. In order to ensure transparency, these components must be reported.

3.6 Recruitment – A clear description for the strategies to be used to achieve the adequate sample size. For example, what pupils or schools are eligible and how will they be recruited?

Data Management

4.1 Data collection methods – Plans for the assessment and collection of baseline outcome and other trial data should be included in the protocol. In addition to the plans, a detailed description of study instruments along with reliability and validity of these assessments. It is important to clearly specify the assessments used, as an advantage of using the prospective cumulative meta-analysis is the standardisation of assessments in the aggregated trials completed in a sequential time period.

4.2 Participant flow diagram – A clear diagram with the number of participants who were randomly assigned, received intended treatment and analysed for primary outcome is required to allow external transparency of the trial.

4.3 Data retention – Plans to promote the retention of participants and how data would be attempted to be collected for participants discontinuing from the protocol.

4.4 Statistical methods – If a separate statistical analysis plan (SAP) is not created and published prior to the trail commencing, the minimum requirements are as follows. Firstly, the protocol should describe the comparisons that will be made between arms of the trial to estimate the effect. Secondly, the statistical methods used in the primary and secondary outcome analysis should be described and

include Hedges g calculation. Thirdly, details of subgroup or additional analyses (e.g. CACE) should be clearly pre-specified.

Implementation and process evaluations

5.1 Research questions – Summary of the research questions to be addressed by the evaluation and process evaluation. If possible, these should link back to the process model.
5.2 Methods – A description of the methods used to address the research questions for the implementation and process evaluation.
5.3 Fidelity – In order to evaluate the programme fidelity, an outline on how this will be assessed should be included in the protocol.

Ethics and dissemination

6.1 Ethics approval – The protocol should include the name and details of the ethics committee responsible for approving the trial. Furthermore, the ethics committee should store the updated protocol if a suitable trial registry is not available.
6.2 Protocol amendments – Plans for communicating important protocol modifications should be included in the protocol. This is important to ensure transparency and enable external reviewers the opportunity to understand why changes have been made to the original protocol.
6.3 Confidentiality – In order to meet data protection standards, a description on how personal information from participants and how this will be collected, stored and maintained should be included.
6.4 Dissemination – Information on the dissemination plans for the researchers and sponsors to communicate the trial findings is a key requirement in a protocol. In educational research, dissemination is a fundamental process which is often poorly delivered. For example, the EEF publish all protocols and technical reports on their website. However, these are not linked to academic research databases so these are unlikely to be found in future systematic reviews.

Appendices

7.1 Ethics materials – Examples of the consent forms, school agreements to participate and other related documents should be available.

7.2 Other related documents – In the development of protocols for aggregated trials in education the use of standardised external assessments may not be possible due to funding restrictions. If researcher made assessments are designed using previous standardised assessment questions (E.g. Test base software), the psychometric properties of the test, including reliability should be completed after the initial pilot. If reliable, copies of the assessments should be provided in the appendix to allow replication of the study.

After the design for a framework in section 8.3 of this chapter was completed, Montgomery et al., (2018) has published the CONSORT – SPI 2018 extension aimed at tackling this very issue for improving RCT reporting. The objective of the study was to develop an extension to CONSORT 2010 for the reporting of RCTs of social and psychological interventions. Thirty-one experts contributed to creating a checklist which extends 9 of the 25 items from CONSORT 2010 to tailor this for social and psychological interventions. This is step in the right direction for the design of RCTs in the social sciences and if adopted by journal editors as a requirement for publication, the quality of the evidence base for interventions will improve.

8.4 Chapter summary

To summarise, the purpose of a protocol is to use this when designing and conducting RCTs in order to increase the rigour and transparency of the research. The introduction of the CONSORT guidelines in healthcare has addressed many of the early concerns raised regarding the transparency and reporting of RCTs (Altman & Dore, 1990; Mahon & Daniel, 1964). These guidelines increased the clarity of the reporting of RCTs in clinical trials, yet the underlining conduct of the study was often weak. Issues regarding deficiencies in clinical trial protocols, such as outcome reporting bias (Dwan et al., 2008) provided serious concerns about the conduct and evidence generated through clinical trials. The creation of SPIRIT 2013 provided guidance to researchers at the initial conception of trials, establishing a framework of good practice and increasing the transparency of RCTs in healthcare. The publication of high quality protocols has strengthened the credibility of trials in healthcare, increasing transparency for external reviewers and other interested stakeholders. The evidence from healthcare literature demonstrates a clear will to make research more rigorous and transparent, from individual researchers to the senior editors of peer reviewed journals. The journal admission policy for the leading academic journals in healthcare are clear in outlining very detailed guidance on research conduct, including checklists such as CONSORT and SPIRIT 2013, with trials registered and protocols published. Furthermore, their appears to be a collective spirit from all stakeholders to learn from past

mistakes, ensuring future research is conducted to the highest standards. However, recent research by Goldacre et al. (2016) into outcome switching in clinical trials highlights that this is still a problem. The research systematically checked every trial published in the top five medical journals to see if they misreported their findings. The study included 67 trials, comparing each clinical trial with its protocol or registry entry. Nine trials were reported without deviation from their protocol, yet the other 58 trials did not report 354 outcomes and 357 new outcomes were included that had not been published in their protocols. On average, each trial in the study reported just 58.2% of its specified outcomes and added 5.3 new outcomes which had not been previously specified. The research demonstrates the importance of publishing trial protocols to allow external reviewers the opportunity to check trial conduct and reporting.

The use of protocols in educational research is very limited, predominantly included if directed by funding institutions such as the EEF. As previously highlighted, the lack of trial protocols creates the same issues encountered in healthcare trials and raises serious concerns regarding the robustness of the evidence. Limited evidence exists in educational research on the issue regarding outcome reporting bias. Pigott et al. (2013) compared reports of educational interventions from dissertations to their published versions. The authors found that non-significant outcomes were 30% more likely to be omitted compared to statistically significant outcomes in the published studies. Unless a systematic review is conducted, it is not possible to ascertain the extent of outcome reporting bias but the lack of trial protocols for the majority of trials makes this impossible to disentangle.

Another point to conclude is the lack of guidance from educational research journals on the conduct and reporting of RCTs in education. Unless the editors collectively agree for the pre-registration of trials, publication of protocols and analysis plans then it will be impossible for external reviewers to assess the validity of the research. The current model of peer review is fundamentally flawed if the goal of educational research involving RCTs is to create robust evidence for systematic reviews and meta-analyses. Until we collectively address the need for developing high quality protocols and the publication of these, the evidence generated by RCTs will continue to be questionable. Yet, if each journal adopted the approach for increasing transparency then particular journals would build a reputation for high quality research.

It is important to attempt to reflect on the potential barriers which currently exist in educational research regarding the creation of protocols and why these are not used more frequently. It is clear that studies funded by the EEF publish high quality protocols for all funded trials, with these accessible to allow transparency for their independent evaluations. Yet, the majority of RCTs published in educational journals do not publish their protocols or analysis plans. Possible reasons for this may

include a lack of requirements from the journal editors for such a requirement, so unless a collective agreement across journals is reached, individual researchers have no incentive to complete these. The lack of a central trial registry or site to publish protocols is also a confounding factor. Yet, with the advances in technology, online repositories for protocol publications should be made available through academic journals.

A simple and cost-effective approach could involve the University ethics boards collectively agreeing to request protocols upon the application for ethics approval. These would be stored or published in a central repository and released to journal editors when studies are peer reviewed before publication. This would allow the peer review process to access the initial protocol and assess potential sources of bias, such as outcome reporting bias.

Finally, this chapter has focused on protocols for predominantly randomised controlled trials, yet the issues are just as significant for other forms of educational research such as observational, longitudinal and case study research. These questions must also be raised across all strands of educational research, making the process transparent and as rigorous as possible to allow confidence in research findings.

Chapter 9: Methodological advantages of using prospective cumulative meta-analyses in education

There has been a worldwide increase in the use of evidence in education through the development of bodies such as the Institute of Educational Sciences' (IES) What Works Clearinghouse (WCC), the Education Department General Administrative Regulations (EDGAR), and the Best Evidence Encyclopaedia (BEE) in the USA. In Scandinavia the Danish Clearinghouse for Educational Research and The Norwegian Knowledge Centre for Education have been established to help evidence inform practice. In Australasia, the New Zealand Best Evidence Syntheses and Australian Teaching and Learning Toolkit in Australia have been set up to inform policy and practice. The UK has also seen a move towards evidence based education through the EPPI Centre and Education Endowment Foundation (EEF). The Sutton Trust-EEF Teaching and Learning Toolkit provides an analysis of evidence to support school leaders with interventions that are based on robust evidence. International networks such as the Campbell Collaboration and Evidence Informed Policy and Practice in Education in Europe (EIPPEE) project have been established, as well teacher-led organisations such as ResearchEd. The key aim of these organisation is to provide evidence for policy makers, school leaders and practitioners with some assurance that evaluated programmes and approaches have met rigorous standards of review. If programmes are selected on the basis of more robust evidence (internal validity), using tested interventions should increase the chance of positive improvements in outcomes if deployed in other schools and contexts.

However, Cheung & Slavin (2015) raise a number of issues regarding the current educational evidence base, specifically relating to certain methodological features being correlated with study effect sizes in meta-analyses and systematic reviews. The impact of methodological features such as the type of publication (published versus unpublished), size of sample (small, N< 250 versus large, N > 250), research design (randomised versus matched) and outcome measurements (experimenter made vs independent) are highlighted by Cheung & Slavin (2015) as methodological factors associated with substantial differences in effect sizes. As stated by Rothstein *et al.* (2005), confidence in meta-analytical and systematic reviews are dependent upon the extent to which the findings are accurate and free from bias. This section will argue that through the use of prospective cumulative meta-analyses in education based on small scale school led aggregated trials, these concerns regarding methodological factors are reduced or eliminated.

9.1 Publication bias

The earliest evidence of publication basis was provided by (Sterling, 1959), whilst reviewing four psychology journals in the 1950s. Sterling found exceptionally high proportions of studies with statistically significant results, with 97% of articles rejecting the null hypothesis. Sterling (1959) suggested that in fields where statistical tests of significance were commonly used, research which yields non-significant results was not published. The study was revisited by Sterling … (1995), reviewing studies in the four psychology journals previously reviewed in the 1950s. The findings suggested very little had changed, with significant positive results in 95% of the articles.

Rosenthal (1980) raised the issue of the file drawer problem, where studies not published are tucked away in file drawers that did not reach the magic .05 level. Rosenthal suggested the failsafe N method to determine if research findings are resistant to the file drawer threat, allowing researchers to establish reasonable boundaries to estimate the degree of damage to the research conclusion that can be done by the file drawer problem.

Through the development of systematic reviews and meta-analyses, the first instance of publication bias was reported by Smith (1980). In a meta-analysis of sex bias in counselling and psychotherapy, not only the magnitude but the direction of effect was different in published and unpublished studies. The effect of published studies was d= .22, demonstrating counsellor bias against females. Unpublished studies was d=-.24, demonstrating counsellor bias in favour of females. Smith concludes that failing to represent unpublished studies in a meta-analysis may produce misleading generalisations.

In medicine, Simes (1986) compared a combination therapy with alkylating agents as a treatment for ovarian cancer. Simes discovered that when registered trials only were included in a meta-analysis, the two treatments under comparison did not appear to affect patient survival. However, including only published trials in the meta-analysis demonstrated a statistically significant benefit for the combination therapy. Simes concluded that assuming the studies registered at the inception represent and unbiased sample of all studies undertaken, the results imply the exclusion of unpublished studies in meta-analyses may lead to erroneously positive results.

Lipsey & Wilson (1993) reviewed 302 meta-analyses relating to psychological, educational and behavioural treatments. They found a ‘strong skew’ towards positive effects, with only six reporting negative effect sizes and few reporting effect sizes around zero. Lipsey and Wilson conducted an analysis of a subset of the data to identify reasons for this positive skew in their data. They analysed 92 meta-analyses and found published studies had mean effect sizes 0.14 standard deviations larger than unpublished studies. It is important to note that the mean effect size estimates for both the

published and unpublished studies fall in a positive range, the published studies were just more positive than the unpublished studies. Logically, this makes sense, as studies with larger effect sizes are more likely to become published.

A recent study by Cheung & Slavin (2015) in education involving 645 studies from 12 reviews of reading, mathematics and science found the overall effect sizes for published articles and unpublished reports were +0.30 and 0.16, respectively. The Q-value (Q_B = 58.47, df = 1, and p,0.00) indicating publication bias in this set of studies. A further study by Polanin et al., (2016) reviewed 383 studies, of which 81 had sufficient information to calculate an effect size. The results indicated that published studies yielded larger effect sizes than unpublished studies (d; = 0.18, 95% confidence interval [0.10, 0.25].

The presence of publication bias in systematic reviews and meta-analyses is a widely known accepted problem in research across a number of disciplines, such as healthcare, psychology and the education (Banks et al., 2012; Dickersin, 1997; Dwan et al., 2008). Assuming that publication bias has been acknowledged widely since late 1980s, and this is still an issue in educational research, the question of why this issue still exists and what are the potential causes are important to ask.

9.1.1 Potential causes of publication bias in research

Torgerson (2006) defines the term publication bias as the tendency for a greater proportion of statistically positive results of experiments to be published and, conversely, a greater proportion of statistically significant or null results not to be published. As previously mentioned, publication bias is not a new concept and a number of reasons have been suggested as a cause.

Firstly, the submission of articles for peer review is an important aspect of investigating study publication bias, with a fundamental distinction needed between whether reports are not published because they are not submitted or they are submitted for peer review but rejected. The initial responsibility lies with researchers as they have a moral and ethical duty to report all findings, regardless of significance to results. Torgerson (2006) suggests researchers have a responsibility to ensure the timely submission of their trials for publication, whatever their results. If it is true that publication of non-significant research is due to researchers not submitting this for peer review, then ethical committees at educational institutions should be an active role ensuring all research is published at their institution. Malički & Marušić (2014) explored the opinions of authors of published clinical trials and Cochrane systematic reviews, demonstrating that researchers are aware of being the main culprits of non-publishing or selective publishing of results from trials. However, the study found

that researchers felt strongly that blame rested not solely with them but with the system that encourages and supports practices that lead to publication bias.

Secondly, as previously mentioned by the opinions of researchers into the reasons for publication bias (Malički & Marušić, 2014), the process of peer review for publication creates the environment for this bias to thrive. It cannot be possible for academic journals to publish all research, however journals should be responsible for publishing non-significant results or studies with minimal effect sizes. If researchers believe that null results have no publication potential, they are more likely to abandon the project and reporting (Franco et al., 2014). If publication of research through the peer review process is due to rejection, journals should provide an alternative database to register these studies.

Potential ambiguity of 'no significant results' can be another potential issue as it is sometimes not clear whether studies remain unpublished because all outcomes are non-significant and those that are published are selectively reported. For instance, after the data is collected the outcomes are modified by the removal of a primary outcome that was not significant. In the social sciences, the lack of trial protocols with outcome and pre-analysis plans often mean this is difficult to identify this type of bias. In clinical trials, a study by Chappell et al., (2005) compared the published version of RCTs in a specialist clinical journal with the original trial protocols. The study found the primary published outcome was exactly the same as in the protocol in only 23% of the trials (6 out of 26 trials), with only 9 trials using the analysis method stated in both the protocol and published version. However, it is important to note that this paper was published at a conference proceeding, without peer review so the results should be viewed with caution. In this instance though, study publication bias and outcome reporting bias overlap.

A recent study in educational research by Pigott et al., (2013) examined outcome reporting bias by comparing the reports of educational interventions from dissertations to their published versions. The authors found that nonsignificant outcomes were 30% more likely to be omitted from a published study than statistically significant ones. The paper found that the majority of authors continued into the career of academia, so it is likely that the pressure of publication could increase the chance of authors omitting nonsignificant outcomes so that papers are more likely to be accepted for publication.

The presence of outcome reporting bias would be expected to influence if an article is published, as significant results and large effect sizes are almost twice as likely to be published (Cheung & Slavin, 2016). As previously mentioned, the lack of trial protocols and outcome and pre-analysis plans prevent any external evaluator disentangling publication and outcome reporting bias in the majority of educational trials. Dwan et al., (2008) explains that the bias from missing outcome data

that may affect a meta-analysis is on two levels: non publication due to a lack of submission or rejection of the study at peer review (a study level problem) and the selective non-reporting of outcomes within published studies on the basis of the results (an outcome level problem). In research literature, the main focus has been on the bias resulting from the lack of submission or rejection of not significant studies. Further research should determine if selective-non reporting of outcomes within published studies is an issue in education, yet this cannot be achieved if educational trials fail to publish protocols, outcomes and pre-analysis plans.

In clinical trials literature, studies have found major discrepancies in the specification of primary outcomes when comparing protocols and published articles (Chan et al., 2004). Overall, the study found primary outcomes were defined in 82 of 102 trials (80%) in either the protocol or published articles. Consequently, 51 of the 82 trials (62%) had major discrepancies between the primary outcomes specified in protocols and those defined in the published article. Table 31 highlights the discrepancies when comparing protocols to the published literature.

Table 31: Proportion of Trials with Major Discrepancies in the Specification of Primary Outcomes When Comparing Protocols and Published Articles

Discrepancy in Published Articles Relative to Protocols	Trials With Discrepancies for ≥1 Primary Outcome, No. (%)*
Primary outcomes specified in protocols (n = 76 trials)	
Any change to protocol-defined primary outcome	40 (53)
Reported as nonprimary in published articles	26 (34)
Omitted from published articles	20 (26)
Primary outcomes specified in published articles (n = 63 trials)	
Any new primary outcome defined in published articles	21 (33)
Changed from nonprimary in protocol to primary in published articles	12 (19)
Not mentioned in protocol	11 (17)
Any discrepancy in primary outcome (n = 82 trials)†	51 (62)

*Trials often defined >1 primary outcome in protocols and published articles.
†Primary outcomes defined in either protocols or published articles.

(Chan et al., 2004)

Table 31 demonstrates clearly discrepancies between protocol and published studies, even when the trial authors developed predesigned protocols. The study authors acknowledge that the adoption of evidence-based reporting guidelines such as the revised CONSORT statement 26 for parallel group trials should help improve the reporting of clinical trials. The current CONSORT guidelines (K. F Schulz et al., 2010) require the following:

- 6a. Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed
- 6b. Any changes to trial outcomes after the trial commenced, with reasons

Regarding 6b, further guidance for authors includes "the need to report all major changes to the protocol, including unplanned changes to eligibility criteria, interventions, examinations, data collection, methods of analysis, and outcomes. Such information is not always reported" (Schulz et al., 2010).

If the social sciences, particularly in education, continue to increase the use of RCTs to inform systematic reviews and meta-analyses, lessons must be learned from the medical profession. The risk of outcome level publication bias is a real threat to the validity of evidence, especially if policy engages with evidence-based practices. If funding is directed towards educational policies and/or practices that have been found to be effective based on meta-analytic research, the issue of the presence and magnitude of publication bias in these reviews must be addressed as a matter of urgency (Banks et al., 2012).

9.1.2 Recommendations from the literature

The report of the first evidence of publication bias in medical literature (Simes, 1986) recommended an international register of all trials as a measure to protect against the effects of publication bias. In order to reduce publication bias in clinical trials, the key recommendation for a number of years has been the creation of clinical trials registers (Dickersin, 1997; Dwan et al., 2008; Kicinski, 2013; Malički & Marušić, 2014; Tonks, 2002; Torgerson, 2006).

The process of setting up trial registries is well underway in healthcare research and this is acknowledged by the profession as a necessary step to reduce the threat of publication bias to the validity of systematic reviews. An important point acknowledged by Dickersin (1997) is that registering all trials at inception does not ensure eventual publication. Nonetheless, registering of trials does make information about all trials available to those performing a systematic review. The process of registering all trials at the inception provides an unbiased sample of all studies undertaken, a key point highlighted by Simes (1986).

The registration of trials in education would be move in the right direction towards addressing the issue of publication bias. After all, the registering of all trials would provide researchers with all results, whether these are significant or not. Although registering trials would be a positive step forward, the evidence from medicine suggests this would not be enough, as outcome bias could still

represent a threat to the validity of systematic and meta-analysis reviews. In addition to the prospective registering of all educational trials at the inception, this is pointless unless the advance publication of detailed protocols with explicit descriptions of outcome and analysis plans are also registered at this stage of the research (Dwan et al., 2008). If alteration between protocol and the published version are unavoidable, readers should be aware of the them (Chappell et al., 2005).

In educational research, journal publications do not require the pre-publication and register of trials. Until a national or international registry of educational trials is developed and academic journals require this as a pre-requisite for publication, publication bias will not go away. However, to avoid publication bias having a detrimental effect on systematic reviews, it is essential that, first, the problem is recognised and second, steps are taken to ameliorate this source of bias (Torgerson, 2006).

Researchers can help minimise the problem of publication bias by carrying out extensive and exhaustive searches of the literature, including unpublished researches in the public domain via electronic databases for grey literature. However, even the most exhaustive searches cannot ensure all the trials have been included. If researchers use the 'file drawer' to store not significant studies, the most sensitive search could not find this. Therefore, as a researcher it becomes important to consider methods of detecting such bias and rectifying the situation.

In educational research literature, Banks et al., (2012) suggests that publication bias analyses, as part of an overall sensitivity analysis, become a requirement for all meta-analytic reviews. The authors reviewed a number of methods for identifying and reporting publications, recommending the following:

1) Contour-enhanced funnel plots
2) The trim and fill analysis
3) The cumulative meta-analysis by precision
4) Begg and Mazumdar's rank correlation
5) Egger's test of the intercept

The authors recommend the discontinuation of the use of the failsafe N to assess publication bias, as the results of the failsafe N were often inconsistent with the results from other publication bias methods. Further evidence (Becker, 2005) supports these findings for the inconsistency of failsafe N to determine and report publication bias.

It is clear that publication bias presents a serious threat for the advancement of educational research, particularly in evidence-based practice. It is clear from the evidence that publication bias

occurs as a result of the failure of research to be submitted for publication and then through the rejection articles at the peer review stage of publication. The file-drawer effect can be addressed using a number of methods as part of a sensitivity analysis, including examples such contour-enhanced funnel plots, trim and fill analysis and cumulative meta-analysis by precision. Furthermore, it is important to distinguish between study level problem of publication bias and outcome level problem of selective non-reporting of outcomes within published studies on the basis of the results. Unless educational practice and journal publications insist on study protocols, outcome and pre-analysis plans then this issue will not be able to be resolved. Through the adoption of the CONSORT guidelines in the use of RCTs in education, this issue might be improved but unless trials are registered centrally and independently from the trial authors, the potential of publication bias and outcome bias will remain a threat to the credibility of meta-analytic evidence-bases.

9.1.3 How can a prospective cumulative meta-analysis prevent publication bias?

The publication of protocols allows a comparison of what was planned with what was actually done, with recent studies in healthcare clinical trials suggesting this comparison should be possible during and after peer review (Hawkey, 2001; Horton, 2006). As discussed in the previous chapter, the introduction of CONSORT and SPIRIT has created a greater transparency and willingness to improve practice in healthcare.

Sadly, in education this is not the case. Unless lessons can be learned and the importance of trial registries, the creation of SPIRIT criteria for the advanced publication of quality trial protocols and publication in journals mandating these requirements for educational interventions using trials, the evidence base will be flawed. However, the development of school led aggregated trials linked to prospective meta-analyses could provide a model of best practice for educational research, ensuring lessons are learned and more importantly implemented, from the healthcare clinical trials.

The first stage in developing school led aggregated trials and reporting these through a prospective cumulative meta-analysis is the development of trial protocols, using the SPIRIT and CONSORT guidelines as the goal standard model. The publication of the trial protocols will improve the transparency in the research, reducing selective reporting bias by publishing the primary and secondary outcomes with pre-analysis plans.

The prospective cumulative meta-analysis will reduce the impact of publication bias, as all schools register in advance and are assigned to cohorts for the delivery of the trials in schools. At each educational institution managing the prospective meta-analysis, the trial registry is updated for each

cohort of schools (school names, number of participants, cohort descriptive data such as SEN, Pupil Premium Status, Gender, and English as additional Language). The initial trial protocol will outline the requirements, with the ethics committee responsible for ensuring protocols are adhered to before granting permission for the trial to commence. As discussed by Simes (1986), the trial registry provides an unbiased sample at the inception of the research for each 'wave' of the database (each time a study is added). Therefore, both positive and negative results are included ensuring the publication bias issue raised earlier with traditional meta-analysis is not an issue. As the prospective meta-analysis eliminates the traditional 'publication' bias in terms of greater reporting of larger effects and statistical significant results, there becomes no need to use methods to detect or rectify this form of bias. Yet, the issue regarding the selective reporting of outcomes can still lead to publication basis but this can also be addressed by learning from the lessons of clinical trials in healthcare.

As mentioned previously, if alterations between the protocol and the published research are unavoidable (Chappell et al., 2005), and in educational research this may well be the case, readers should be aware of these through clear reporting in the published document. In the publication of research using prospective cumulative meta-analyses, all protocols and cohort registries of schools are published with the dates of the ethical reviews for each wave of trials. This would allow a transparency in the data, allowing independent analysis for the comparison of protocols for what was planned with what actually happened. The greater transparency will ensure outcome reporting bias is greatly reduced, which is currently not possible in the majority of educational research at this present moment in time.

9.2 Size of sample

In has often been acknowledged in systematic review literature that studies with small sample sizes generally tend to have larger effect sizes than studies with larger sample sizes. In education, the phenomenon has been identified in a number of studies (Cheung & Slavin, 2015; Slavin, Lake, & Groff, 2009; Slavin, Cheung, Groff, Lake, 2008) with a substantial negative relationship found between sample size and effect size. In medicine and psychology, a number of studies have also highlighted the potential issue of small size and effect sizes (Ioannidis et al., 1998; Kühberger et al., 2014) with a number of reasons suggested to explain the apparent negative relationship between sample size and effect size. For this reason it becomes important to understand why sample size has an impact on effect sizes and how a prospective cumulative meta-analysis using small scale aggregated trials can minimise this impact.

9.2.1 What literature exists in education for influence of sample size and effect sizes?

In education, a meta-analysis of the effects of hypermedia on students' achievement, Liao (1999) found a grand mean effect size of +0.41 for the 46 studies included in the meta-analysis in favour of hypermedia instruction. However, Liao acknowledges approximately 65% of the studies included had samples with less than 80 participants (mean effect size +0.60), with larger sample size studies having a much smaller effect size (mean effect size +0.03). The author suggests that the effects of hypermedia on students' attainment may only work for small to medium samples. It is important to recognise a limitation of this study is the high proportion of small sample studies with positive effect sizes, as these will be under powered, with many of these small trials possibly including false positive results and be particularly threatened by publication bias.

Moran et al., (2008) discovered that studies with smaller sample sizes were much more likely to achieve substantial effects than those with larger sample sizes. The authors argue that this may be due to researchers maintaining high degrees of fidelity to treatment in smaller studies because of greater manageable prospects. However, Moran et al. explain that this is counter-intuitive and puzzling due to the increase in statistical power in larger studies. Yet, it should be argued that this is to be expected, as smaller samples require larger effect sizes to become statistical significant. This will be discussed further in this section, as other potential moderators such as publication bias, super-realisation and small study variation can explain this phenomenon.

Slavin & Smith (2009) investigated the relationship between sample size and effect sizes in systematic reviews in education, analysing data from 185 studies of elementary and secondary mathematics programmes. The study included published and unpublished sources, with no restriction on sample size but the inclusion criteria required studies to meet the standards set out in the Best Evidence Encyclopaedia. Sample sizes ranged from 30 to 40,000 so the authors recoded that data into eight categories to reduce the influence of extreme scores. The results of the study found the overall correlation was -.28 ($p < .001$), with smaller studies with sample sizes below 50 averaging +0.44 effect size and studies with samples greater than 2000 having average effect sizes of +0.09. A potential limitation of the study was the number of studies included in the review, however further evidence from a recent study by Cheung & Slavin (2015) provides further evidence to support their conclusions. This study completed a similar comparison but with a larger set of studies. The study included 335 studies with a sample size less than 250 and 310 studies with greater than 250 participants. The authors found studies with smaller sample sizes (ES= +0.30) produced almost twice the effect sizes of studies with larger sample sizes (ES= +0.16). A statistical significant difference was found between large and small studies ($Q_B = 55.28$, df=1, and $p<0.00$).

In the area of peer tutoring in education, research from meta-analyses predominantly finds positive effect sizes for attainment (Cohen et al., 1982; Kim Chau Leung, 2015; Ritter, Barnett, Denny, & Albin, 2009; Rohrbeck et al., 2003). The EEF commissioned two large scale randomised trials involving nearly 10,000 children in UK schools (Lloyd, et al., 2015a; Lloyd, et al., 2015b). The studies found very little impact, with shared maths effect size (0.01 and 0.03) and peer tutoring in secondary schools ES (-0.02 and -0.06). Therefore, when the intervention was scaled and evaluated independently from the developers, the effect sizes were minimal. Consequently, we must try to understand why this phenomena occurs and the possible reasons why small scale trials often report larger effect sizes than larger studies.

In educational settings, Slavin & Smith (2009) acknowledge 'super-realization' as the main source of small study effects. The term was introduced by Cronbach et al., (1980) to refer to the fact that in small scale studies, researchers are able to monitor the quality of implementation and provide additional assistance creating unrealistic conditions. When the study is scaled up, it becomes impossible to replicate the trial. The internal validity of the study is not affected by super-realisation, yet it can have significant impact on external validity. In can be argued that in educational intervention research, the external validity should be equally important as the internal validity. Unless an intervention can be replicated and scaled up to be delivered in classrooms, studies with large effects in small studies which cannot produce a meaningful effect size once scaled up should be regarded with caution in education.

A second reason suggested for small study effects is the presence of publication bias within the sample of literature included in the systematic reviews and meta-analyses. Cheung & Slavin (2015) argue that small studies may appear to have high effect sizes because their limited statistical power requires high effect sizes to reach statistical significance. As previously mentioned in the publication bias section, these are more likely to be published in peer reviewed journals compared to small studies with small or negative effect sizes, creating a biased sample.

Kyaergard, Villumsen & Gluud (2001) offer a further confounding moderator when attempting to explain why smaller studies tend to provide larger effect sizes. The authors suggest a weakness in methodological quality in smaller studies as a potential source of bias, as smaller studies are often pilot studies. These pilot and small-scale studies usually have sample sizes which are too small to use a cluster design and analysis which controls for nesting within clusters. Sometimes in the reporting this can be 'glossed over' through citing it as a limitation and yet the results are still published and they can then find their way into a meta-analysis. Consequently, the intervention is often delivered by the teacher and, thus, at class level. When the analysis is done, the effect size is larger because the proper design and analysis methods have not been used. Even if a protocol has been used, this will

still contribute to publication bias. In educational research, this could provide a possible explanation as pilots often include inferior measurements (researcher made assessments) and poor controls for selection bias, such as blinding in the allocation, assessment and analysis. Hróbjartsson et al., (2014) found that a lack of blinded outcome assessors in randomized trials with subjective time-to-event outcomes causes high risk of observer bias. Non-blinded outcome assessors typically favour the experimental intervention, exaggerating the hazard ratio by an average of approximately 27%. Furthermore, Ainsworth et al., (2014) found in a study comparing the blinding of primary and secondary outcomes that in their particular trial, the difference between the primary and secondary outcomes was likely to have been due to lack of blinding of testers.

Finally, effect sizes in smaller studies are likely to be more variable than in larger studies, especially in educational research when students are clustered in a small number of classes or schools. For example, if students were randomly assigned to classes with only three experimental and three control teachers, teacher and class effects are confounded with treatment effects. This could be positive or negative, and a single good or poor teacher could determine outcomes far more than the treatment effects. Therefore, small studies are likely to contain a disproportionate number of very positive or negative effect sizes. In the published research literature, these do not balance out due to the publication of mainly positive results.

It is clear from the literature that the issue of small study effects can have a significant potential to undermine the validity and provide potentially biased evidence when considering the effectiveness of interventions in education. So how would the development of prospective cumulative meta-analyses in education using small scale aggregated trials minimise these issues?

9.2.2 How can a prospective cumulative meta-analysis prevent the impact of small sample sizes on effect sizes in education?

The methodological advantages of using a prospective cumulative meta-analysis of school based aggregated randomised controlled trials ensure the issues identified earlier relating to the effect of sample size on effect sizes are minimal. Firstly, the introduction of pre-protocols and trial registration ensures all results, either positive or negative are included in the analysis thus creating an unbiased sample.

Secondly, as smaller studies require larger effect sizes to produce statistical significance compared to larger studies, a cumulative meta-analysis again eliminates this source of bias. The

aggregation of data ensures the sample size increases continuously, as demonstrated in the early cumulative meta-analyses used in clinical trials (Lau et al., 1995).

Thirdly, the inclusion of pre-protocols counters the issue of poor methodological quality that is often associated with small studies (Kyaergard, Villumsen & Gluud, 2001). As smaller studies are often pilot studies, these tend to include inferior measurements and poor controls for selection bias. The lack of methodological quality can result in greater study to study variability and produce bias toward positive effects. It is important to note that Kyaergard (2001) refers to clinical trials in medicine, however the argument can be applied to the education as a potential threat to the quality of small scale studies.

As previously mentioned, Slavin & Smith (2009) acknowledge 'super-realization' as the main source of small study effects in educational research. The additional support often provided by program developers and researchers in small scale studies cannot be replicated when the intervention is scaled up for large trials. This phenomenon has been highlighted by Wiglesworth (2016) as studies coded as 'efficacy' will show larger effect sizes compared to studies coded as 'effectiveness'. The results of the meta-analysis indicate a trend towards greater effects when additional support, training, staff or resources are provided. As small studies will tend to have close support from researchers, it seems logical that this could have a positive influence on the effect size. In prospective cumulative meta-analyses using school led small scale trials, the impact of the researcher would be less direct and minimal as the implementation is managed by school nominated staff.

Cumulative meta-analyses created through small scale aggregated trials will also reduce the impact of the variability between traditional small scale and large scale studies. In educational research, students are generally clustered in a small number of classes or schools, which may add class or school effects in one direction or another. Through the aggregation of small scale trials, the sample size increases after each addition of the previous study. The impact of potential bias from class or school effects should reduce as the sample size increases, as it does with large scale studies, with the forest plots in the cumulative meta-analysis demonstrating smaller confidence intervals. As the sample size increases through the aggregation of trail data, the effect size observed is more likely to be closer to the true effect of the intervention.

In traditional meta-analyses, recommendations for addressing the issue if small sample sizes include the weighting of effect sizes by their sample sizes and the addition of minimum sample sizes in the inclusion criteria (Slavin & Smith, 2009). A recent recommendation by Cheung & Slavin (2015) suggest researchers should be encouraged to pool similar studies to build up a sample size over time. However, as these small studies were not prospectively planned, issues regarding the outcome measures, methodological quality and length of the intervention are still potential issues. It becomes

clear that the prospective development of small scale aggregated trails linked to a cumulative meta-analysis would develop a robust evidence base for evaluating educational interventions.

9.3 Research design

In the development of future prospective cumulative meta-analyses, it is important to consider whether the research design chosen for an aggregated trial could potential impact on the effect size reported. At present, the type of research design (randomised experimental or quasi-experimental) has produced mixed results when considering their impact on effect sizes. For this reason it becomes clear that we need to understand why research design has an impact on effect sizes and how a prospective cumulative meta-analysis using small scale aggregated trials can minimise this impact.

9.3.1 What literature exists in education for influence of research design and effect sizes?

A recent meta-analysis of the effects of attributes on student academic performance found no difference in the meta-regression in the effect between interventions in which students were assigned randomly or not to experimental and control groups (de Boer et al., 2014). Previous research by Heinsman & Shadish (1996) concluded that if randomised and non-randomised experiments were equally well designed and executed, these would yield a similar effect size. A similar conclusion was found in a systematic review for the impact of computer technology on mathematics education in K12 classrooms (Li & Ma, 2014), with true experimental and quasi-experimental producing similar effect sizes. Melby-Lervåg & Hulme (2013) conducted a meta-analytic review of working memory, coding studies as randomised or non-random assignment. No significant differences between these designs were observed in the study.

In contrast to these study findings, Cheung & Slavin (2015) categorised 196 studies as randomised and 449 as well-matched quasi experimental studies. The authors found the average effect size for randomised studies as +0.16, compared to +0.23 for quasi-experimental studies. Cheung & Slavin suggest the increase could be due to selective factors in favour of treatment groups when random assignment is not used. Even if schools are matched on quantitative factors such as free school meals, performance pre-tests, English as additional language, it is impossible to balance out unseen characteristics. For example, if the schools in the treatment have volunteered to participate in the study, these schools may have stronger leadership, more innovative staff who may be more condiment in their abilities. Only random assignment to the treatment and control could balance these

unseen characteristics. Interestingly, when research design is combined with sample size (more or less than 250), small matched studies (n=229) produce higher effect sizes (+0.33) compared to large randomised trials (n=90, ES=+0.12).

Previously, Slavin, Hanley, Lake, & Hanley (2012) reported in a Best Evidence Encyclopaedia review of effective science programmes that effect size increased as experimental design moved from matched and clustered randomised controlled trials to quasi-experimental design. A recent study by Zeneli, Thurston, & Roseth (2016) collaborates the evidence from Cheung & Slavin (2015), suggesting that the stricter the research design, the smaller the effect size would be observed. However, this study highlights a limitation as the small number of quasi-experimental design studies in the review did not allow a sensible comparison with other designs.

9.3.2 How can a prospective cumulative meta-analysis prevent the impact of research design on effect sizes in education?

The evidence of the potential impact of research design on effect sizes is mixed, however the finding from Cheung & Slavin (2015) comparing small matched studies with large randomised controlled trials has importance for policy makers and educational researchers. The research found small matched studies produced effect sizes (+0.33) significantly larger than large randomised trials (+0.12). The significance of this finding cannot be overlooked, as organisations such as the WWC give its highest rating of confidence to only well implemented Randomised Controlled Trials. In the United Kingdom, the EEF uses independent peer reviewers to allocate a security rating for trials, with well conducted Randomised Controlled Trials able to achieve the maximum five padlocks. However, power and attrition, as well as baseline equivalence and internal threats to validity can reduce the security rating of a trial.

The purpose of randomisation is to safeguard against potential selection bias, as this is impossible ensure treatment and control groups are balanced on unseen characteristics, such as teacher motivation or if pupils have access to technology outside of school. As quasi-experimental studies are often less expensive and more feasible in educational settings, the use of these when comparing effect sizes with large randomised trials should be interpreted with caution. As previously stated recommended by Cheung & Slavin (2015), the cumulative pooling of small randomised trials would increase the sample size and power of smaller studies. By virtue of the nature of aggregated trials, the pre-protocol creates a homogenous study with all the smaller trials following strict guidance.

Pre-protocols ensure the sample characteristics (age of pupils, ability, SEN status[20], EAL[21], FSM[22]), randomisation procedure, implementation procedures for the intervention, outcome measures and analysis methods are standardised. Traditional meta-analyses often have to consider heterogeneous studies using different study designs, outcome measures and pupil characteristics. Yet, the prospective cumulative meta-analysis ensures that each cohort of schools in the smaller trials uses the same research design, as well as other key methodological features. Therefore, researchers involved in designing the pre-protocols should attempt to use true experimental designs before considering quasi-experimental, justifying their reasons if they choose to use matched rather than random assignment. If matched designs are selected, researchers should use every possible means to avoid selection bias and make sure treatment and control groups are comparable.

9.4 Outcome measures

In educational research, the issue of researcher made versus standardised assessments is often a dilemma researchers' encounter when designing a trial. Logically, using standardised tests in any trial should reduce potential sources of bias, especially if these are administrated blinded and marked independently of the evaluator. However, for practical reasons many small scale studies develop researcher made assessments as a measurement instrument for a trial. Ideally, researchers would conduct a RASCH and Alpha Cronbach analysis to establish the reliability of the test if the decision is taken to use non-standardised assessments. Consequently, using researcher made assessments does decrease the internal validity of any trial, but how much does this potentially impact on the effect size?

9.4.1 What literature exists in education for influence of researcher made versus standardised assessments and effect sizes?

Scammacca et al. (2007) conducted a meta-analysis on reading interventions for struggling readers from 1980 to 2004. The meta-analysis included a total of 33 studies, finding an overall mean effect size of 0.95 across all types of reading interventions and outcome measures. However, when the studies were analysed based on the type of assessment measure used, the mean effect size for standardised, norm-referenced measures was 0.42. The report has a number of potential limitations, including small sample sizes (studies with 13 participants) and only included published studies,

[20] SEN status refers to special educational needs.
[21] EAL refers to English as additional language
[22] FSM refers to Free School Meal status, an indicator of social deprivation

excluding unpublished reports, dissertations and theses. Consequently, publication bias could and would most probably be likely within this study due to the exclusion and inclusion of small sample size trials. These effects are also reported by Edmonds et al. (2009) in a meta-analysis of reading interventions and effects on reading. The mean effect size of researcher developed assessments (ES=+1.19) were significantly larger than the studies using standardised assessments (+0.47). As with the previous meta-analysis for struggling readers by Scammacca et al. (2007), the study excludes unpublished literature and includes a number of studies with a very short implementation time. Therefore, publication bias could present a threat to the validity of the findings and interventions delivered over days or a few weeks could be influenced by the Hawthorne effect. In the above mentioned studies, researcher made assessments tend to report effect sizes twice the size of trials using standardised assessments. As these studies focused on reading interventions so it is important to establish if the effect of researcher made assessments occurs in other subjects in education.

Li & Ma (2014) conducted a meta-analysis of the effects of computer technology on school student's mathematics learning. The effect size of trials using researcher made assessments (ES=+0.86) was higher than trials using standardised assessments (ES=+0.57). Cheung & Slavin (2015) included mathematics, Science and reading programs in their meta-analysis on how methodological features affect effect sizes in education. The effect sizes for researcher made assessments (ES=+0.40) were twice the size of standardised assessments (ES=+0.20), collaborating the findings from the previous studies mentioned.

Finally, Scammacca et al. (2013) updated the previous study (Scammacca et al. 2007) to include studies from 1980 to 2011. The overall mean effect size (ES=+0.49) had reduced from the previous study (ES=+0.95), with the mean effect for studies using standardised tests (ES=+0.21) smaller than the reported effect size (ES=+0.42) in 2007. The suggested reasons for the decline in the effect size include the increased use of standardised assessments, more rigorous research designs and improvements in the "business-as-usual" instruction. However, the mean effect size of the reading interventions (ES=+0.49) is still twice the size of standardised assessments (ES=+0.21). Again, methodologically the study was flawed, with only published studies in academic journals considered for inclusion in the study. Therefore, publication bias cannot be ruled out as a threat to the conclusions for the effectiveness of the interventions, yet this should not influence the findings comparing researcher made and standardised test inclusion. The evidence presented from the studies above clearly shows that researcher made assessments in trials creates a potential source of bias, resulting in effect sizes twice the size of trials using standardised measurements.

9.4.2 How can a prospective cumulative meta-analysis reduce the impact of researcher made assessments on effect sizes in education?

It is clear that in the design of the pre-protocol for a prospective cumulative meta-analysis, a threat to the internal validity of the trial would be the inclusion of researcher made assessments. Whenever possible, standardised assessments should be used to measure primary and secondary outcomes. Yet, in certain circumstances it might not be feasible to implement standardised assessments due to financial restrictions or fitness for purpose[23]. If researcher made assessments are used, the reasons for inclusion must be transparent in the pre-protocol and acknowledged as a limitation in the full written report. In addition to this, it should be expected that the internal reliability and other appropriate psychometric properties of the researcher made assessments are reported.

9.5 Conclusions

As educational policy makers and school leaders move towards an evidence-based system, it is crucial that the evidence presented is of the highest quality, accurate and free from bias. Randomised Controlled Trials (RCTs) are considered by the WWC and EEF as the 'gold standard', as a "well-implemented RCT is an important tool for finding an unbiased estimate of the causal effect of deliberate interventions" (Ginsburg & Smith, 2016). The strength of a well implemented RCT lies in the strong internal validity, with the ability to combine studies through systematic reviews or meta-analyses. It is imperative that the conclusions and effect sizes established from RCTs and meta-analyses are robust and trustworthy, particularly when funding is limited due to fragile economic conditions on a global scale. Yet, as the evidence presented in this chapter shows a number of methodological flaws are common in current research practices.

Firstly, the issue of publication bias presents a threat to all systematic reviews and meta-analyses in education. In a recent review of 187 systematic reviews and meta-analyses included in the EEF teaching and learning toolkit, only 35% of studies acknowledge publication bias a potential threat to conclusions. Furthermore, from the 35% of studies testing for publication bias, 50% use inappropriate methods such as failsafe N or Orwins's failsafe. The evidence has been consistent over a sustained period of time, whereby published studies tend to have effect sizes twice the size of unpublished studies (Cheung & Slavin, 2016; Lipsey & Wilson, 1993; Polanin et al., 2015; Rosenthal,

[23] Fitness for purpose means that standardised assessments are not appropriate for measuring the construct in the trial.

1980; Simes, 1986). Through the nature of a prospective cumulative meta-analysis based on school led aggregated trials, all trials are pre-registered with results published, whether these are positive or negative. The cumulative meta-analysis creates an unbiased sample, eliminating the threat posed by publication bias.

Secondly, outcome selection bias cannot be determined in the vast majority of educational trials due to the lack of trial registration and pre-published protocols. Examples from clinical trials (Chan et al., 2004) and education (Pigott et al., 2013) demonstrate how outcome selection bias can threaten the validity of research. If research adopts the methodology of a prospective cumulative meta-analysis, the publication of pre-trial protocols would allow independent researchers the opportunity to examine differences between these and publish studies. Through greater transparency, the threat of outcome selection bias would be minimised.

Thirdly, the evidence suggests that the sample size of studies can influence the effect sizes, with smaller studies (n<250) having twice the effect size as larger studies (n>250). (Cheung & Slavin, 2015; Liao, 1999; Moran et al., 2008; Slavin & Smith, 2009). Suggested reasons include 'super-realisation', whereby researchers are able to monitor the quality of implementation and provide unrealistic assistance in the delivery of the programmes. Once scaled up, it is impossible to replicate these conditions. A common criticism of single RCTs identified by Ginsburg & Smith (2016) is that a single well implemented RCT has very strong internal validity, it is carried out on a particular sample, place and time in only one setting. The external validity of RCTs is often questioned. A recommendation proposed by Shadish, Cook & Campbell (2002) is to develop a meta-analysis of multiple trials to help establish external validity. The development of a prospective cumulative meta-analysis would provide the solution to this issue, allowing small scale RCTs with strong internal validity to be run across schools in a country such as the UK. As the sample size increases, due to cumulative data collection, the confidence in the actual effect of the intervention can be increased.

Fourthly, the decision to include researcher made or standardised assessments can also influence the internal validity of any trial (Cheung & Slavin, 2015; Edmonds et al., 2009; Li & Ma, 2014; Scammacca et al., 2013; Scammacca et al., 2007), even in the development of a prospective cumulative meta-analysis. Therefore, standardised assessments should be used whenever possible with researchers required to justify the inclusion of researcher made assessments at the pre-protocol stage of a trial. The researchers would then be required to acknowledge this limitation in the published report, as a threat to the validity of their findings.

In addition to the methodological features mentioned above, the stage of trial development can also impact on the size of the effect size reported. As previously mentioned in Chapter 2, an

intervention should be tested at several stages between its conception and dissemination in practice (Greenburg et al., 2005). A particular focus of RCTs occurs in either the efficacy stage or effective stage of trial development. It is important to make the distinction between efficacy and effectiveness, as research evidence suggests that stage can influence the size of the effect size in trials.

Wiglesworth et al. (2016) define an efficacy study as being conducted under highly controlled and optimal conditions in order to maximise outcomes. Conversely, an effective study is conducted in multiple natural settings, using only the resources that would be available under normal circumstances. In a meta-analysis on the impact of trial stage involving 89 studies (Wiglesworth, 2015) found greater effects under efficacy conditions in all but one outcome variable. The results agree with the logic that if more support is provided, then implementation should be higher and a higher effect size should be found. As previously explained in chapter 1, the development of a prospective cumulative meta-analysis linked to small scale school led aggregated trials would occur in the dissemination stage of the trial cycle. Therefore, a prospective cumulative meta-analysis should not be used in the efficacy stage where researcher involvement is high. The ultimate aim of prospective cumulative meta-analysis is to collect data on the actual effect that the intervention is having in schools, reporting the findings back to researchers to monitor the impact over a sustained period of time involving a large sample of schools. However, a potential issue raised in the use of prospective cumulative meta-analyses in clinical trials is the criterion for stopping the data collection. Chapter 10 will outline potential limitations and how these could be resolved in the development of cumulative meta-analyses in education.

In sum, the development of a methodology for a prospective cumulative meta-analysis linked to small scale RCTs in schools provides practical solutions to a number of methodological issues raised by Cheung & Slavin (2015). As stated earlier by Rothstein *et al.* (2005), confidence in meta-analytical and systematic reviews are dependent upon the extent to which the findings are accurate and free from bias. Through the use of small scale RCTs with high internal validity, the prospective registering of schools over a period of time allows the accumulation of data from a wide range of schools across a country such as the UK. Consequently, the previous concerns regarding the external validity of RCTs is reduced as the sample size increases over time. The development of the methodology of a prospective cumulative meta-analysis may provide researchers, policy makers, school leaders and teachers with an up to date 'what works now in schools' overview for interventions and identify possible groups where interventions are successful, or just as importantly not having the desired impact.

Chapter 10: Discussion, limitations and conclusions

This chapter discusses the findings from the primary research studies, lessons learned for implementing small scale RCTs in the development of a prospective cumulative meta-analysis, and makes the case for why social sciences should adopt a standardised approach to evidence-based education and how the research community could involve teachers in pragmatic small scale RCTs to deliver higher quality in school evaluations of what works. The chapter then progresses on to review the study limitations; discussing these in terms of theoretical, methodological and implementation issues before suggesting key recommendations for policy and practice. Finally, the conclusion outlines the importance of the thesis to the current evidence base in education and further areas for future studies.

10.1 Discussion on the evidence from the primary research studies

The study describes and reports two independent primary research studies which involved the use of online learning for cross-age peer tuition and small group teaching. The findings are not intended to be generalisable due the small sample size, and the primary purpose of the research involved the development of the methodology to use small scale aggregated trials across a sequential time period using a specific protocol linked to a prospective cumulative meta-analysis (PCM). Consequently, the findings from the individual research questions are limited within the scope of the initial research. Nevertheless, these findings are important as both studies contribute to the evidence for online learning in the context of peer tutoring and small group tuition.

10.1.1 Pilot efficacy and feasibility study for small scale aggregated trials investigating the effectiveness of online cross-age peer tuition group sizes across the transition between primary and secondary schools

The transition between primary and secondary schools in the UK has been highlighted by Ofsted as a key area for improvement for schools (Ofsted, 2015). The report 'KS3: The wasted years' highlights the lack of an academic transition for the more able and talented pupils during the transition from primary to secondary schools. Harrison (2015) successfully implemented a feasibility study for online cross-age peer tuition across the transition boundary between primary and secondary schools using an active control (1:1) to compare the relative effectiveness of 1:2 and 1:4 online group peer tuition in mathematics. The current study follows on from the initial research to implement the intervention

on a larger scale using a protocol and small-scale aggregated trials to develop a PCM evidence base for the intervention.

The research evidence on the impact of peer tuition is fairly robust, with a number of meta-analyses providing evidence that peer tutoring has a positive impact on tutee achievement (Cohen, Kulik, & Kulik, 1982; Cook et al., 1985; Ginsburg-Block, Rohrbeck, & Fantuzzo, 2006; Leung, 2015).

The research literature for the evidence base on the impact of online peer tuition on academic attainment is much weaker, with many studies using attitudinal responses through surveys rather than focusing on the impact on attainment. The current study uses synchronous communication using real time communication and collaboration between the tutors and learners. The research evidence on the effectiveness of synchronous versus asynchronous delivery on learning is variable (Hrastinski, 2006; Rockinson-Szapkiw, 2009; Schwier & Balbar, 2002) due to the research designs used in the studies.

Tsuei (2012) used a quasi-experimental design for synchronous online peer tutoring in mathematics for pupils aged 10-11 years and found the experimental group had significant learning gains compared to the control. However, the sample size of 88 children was insufficiently powered to generalise these findings.

The current study contributes to the evidence base as it demonstrates possibility of using robust research designs to evaluate the effectiveness of an online intervention. The use of an active control in the design ensured all pupils participated in the intervention and allowed a comparison for the effectiveness of the group sizes 1:2 and 1:4 with 1:1 peer learning. The PCM provides an aggregated effect size +0.05 (-0.44, 0.55) for 1:2 compared to 1:1 and +0.06 (-0.46, 0.58) for 1:4 compared to the 1:1 active control. It is acknowledged that the sample does not allow these finding to be generalised outside the current sample population, however the PCM in figure 10. demonstrates a fairly stable effect size for each cohort for the group sizes.

Furthermore, the study has demonstrated the potential to aggregate data from small scale trials to create a PCM as an evidence base for a specific intervention. In relation to the evidence-base for online learning, I believe this is the first instance for the use of a PCM linked to small scale trials.

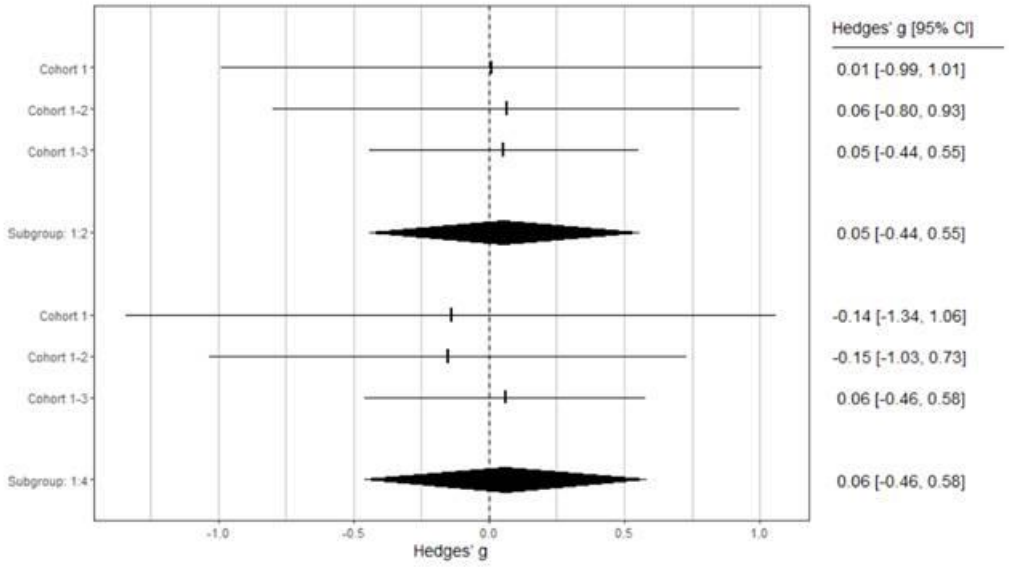


Figure 27: PCM for the aggregated trials for online peer tutoring group sizes 1:2 and 1:4 compared to the active control 1:1.

Finally, the study has demonstrated that online peer tutoring can be used as an academic intervention to support the transition between primary and secondary schools.

10.1.2 Online small group teaching

As previously discussed in earlier chapters, the evidence base for small group teaching and online group learning is limited. The trial focused on online small group learning and adds to the evidence-base interventions that improve Mathematics in secondary school children. However, as the study was stopped after the first cohort of schools, the sample size is not sufficient to generalise and should be treated with caution. The main objective of this RCT was to test the gains in mathematics ability against those of pupils in the control group. The criteria of a trial were maintained throughout and no schools dropped out of the evaluation, the TUTE intervention was delivered over a 5-week period, the pupils in the control did not have access to the TUTE system and the attrition rate was 5%. The overall finding suggests that TUTE was an effective mathematics intervention with the effect size of around +0.44. However, at school level the effect size ranged from -0.57 to +0.74, demonstrating the need for caution with such a small sample size.

10.2 Discussion and personal reflection on the implementation of small-scale trials linked to the aggregation of data for a prospective cumulative meta-analysis

The implication that evidence-based education involves interventions, whether these are implementing policy, providing advice or specific educational interventions, imply that the purpose of the intervention is to bring about change. For example, a classroom teacher would not look to implement a new peer tutoring intervention in mathematics if this was not to bring about change, either academically to raise attainment or to develop the social skills and confidence of the peer tutors. The definition provided by Coe et al., (2000) includes the term 'good evidence', implying a hierarchy whereby the type of research design implies a level of confidence that the intervention was actually responsible for any changes in outcome. Therefore the research question involving the implementation of the intervention or change is causal, as the intervention (X) is intended to bring about a change (Y) and the researcher must ensure that no other plausible explanations can explain this change.

As previously discussed in Chapter 2, Biesta (2007) and Hammersley (1997) argue against the idea that education is a causal process but a process of symbolic or symbolically mediated interaction. Morrison (2001) uses complexity theory to replace the emphasis on causality, stating that small changes in initial conditions can produce massive and unpredictable changes in outcomes for learners. I have argued that fundamentally these positions are not coherent, as this suggests that there could never be an intervention whose effects are predictably beneficial. Do we really believe that research could not inform practice?

Evidence-based education does involve the assumption that when an intervention or change is implemented, a causal mechanism is inferred, but it is crucial to understand the counterfactual or comparison being made. Higgins (2017) explains that through different kinds of comparison, such as different research designs, we are able to provide stronger or weaker arguments about the robustness of any causal claims relating to the effectiveness of an intervention. In the example relating to a teacher implementing a peer tutoring intervention in mathematics, the teacher could use numerous research designs to determine if the intervention had a positive impact on attainment. In one instance, the teacher could select 30 Year 9 pupils from their top set mathematics class to peer teach 30 year 7 pupils for 6 weeks on the topic of data handling. Using a pre-test and post-test the teacher is able to show that all the pupils who participated increased their test scores over this period and they could present robust data to their line manager on the positive benefits of the intervention.

A single group design such as this compares the outcomes of the same pupils before and after the change. As no comparator group is used in this design we can make minimal causal inference as we do not know how much the pupils would have improved without the intervention. Numerous

threats to the internal validity of any research can provide an alternative plausible explanation for the increase in attainment such as selection bias, regression to the mean, maturation and temporal changes, instrumentation threats, attrition and social interaction threats.

Alternatively, the teacher could use either a quasi-experimental or randomised controlled trial design. In this particular example, the teacher would compare the average outcomes from the groups allocated by random assignment with a group that does not experience the change, creating equivalent groups so that the known and unknown variables balance when the sample size is sufficient. In this design, the threats to internal validity can be minimised for selection bias, temporal changes, regression to the mean and instrumentation threats. The design still has limitations as attrition and social interaction threats can still impact the internal validity. However, when comparing these two designs it is clear that the use of a control group and randomisation provides a stronger argument for the robustness of any causal claims relating to the effectiveness of the peer tutoring intervention. Yet, how many teachers or school leaders will know the difference between these two designs, and the methodological limitations involved with each approach?

In schools, many teachers are involved in delivering interventions on a daily basis inside and outside of the classroom to help close the gap in attainment with learners who are underachieving. However, teachers are not trained in research design and do not necessarily understand evidence-based education (Cain, 2015; Cordingley, 2015). As a previous teacher and middle leader in a secondary school, research and evidence was not a priority in the day to day teaching role, as marking, planning and behaviour management were rightly at the forefront of the to do list. Consequently, as an educator the importance of understanding research only became apparent during my studies for an MA in Research Methods before commencing the PhD research underpinning this thesis. The research evidence reflects this disconnect between practitioners and policy makers with the research producers (Lysenko et al., 2014; Levin, 2013; Sharples, 2013; Shephard, 2007), in that research is conducted by one set of people then shared with another. In educational research, a number of theoretical models for bridging the gap have been suggested, yet are we fundamentally missing the point in the research cycle before we try to disseminate the findings with practitioners and policymakers? Should research be co-produced by researchers and practitioners working together through every of the process? Would a symbiotic relationship achieve a greater success than a dissemination approach?

Reflecting on these questions and personal experience when transitioning from a practitioner to a researcher, as a practitioner I did not have the knowledge to design research or fully appreciate how to conduct research in a school setting. Even if as a teacher I had completed CPD (continued professional development), the pressures involved in the role of a classroom teacher would generally

mean that I would file the handouts away and rarely look at these again. I understand these are personal reflections and cannot be generalisable to the wider teaching population, yet Lysenko et al. (2014) found that when teachers were asked to position their use of research, the respondents positioned their use at the low end of the scale, between 'never' and 'once or twice' during the last year. These findings support earlier studies in the low use of research by school practitioners (Cousins & Walker, 2000; Williams & Coles, 2007).

Therefore, when considering the research cycle as outlined by Gorard (2013) we should involve practitioners throughout the whole cycle and not just at the dissemination stage of research (Figure 6).

Practitioners may not have the skills and the knowledge to design or evaluate research, yet they do have the key knowledge of implementing interventions in school environments. Research has demonstrated that school-led trials for Accelerated Reader programmes and phonic interventions (Gorard, Siddiqui & See., 2015) involving researchers and practitioners can produce robust evidence on the impact of educational interventions. Siddiqui et al. (2015) highlight that through schools running their own trials, they found a lower cost, easier permission to innovate and less drop-out.

The previous studies aggregated data for the randomised controlled trial at one specific point in time, whereas in this thesis the context of an aggregated trial is defined as the aggregation of separate small scale randomised controlled trials following a pre-defined protocol over a sequential time period. The difference is important in the context of the research cycle, as aggregated trials can be used either at the phase 3 for the feasibility studies, phase 4 and 5 for prototyping and trialling and field studies and instrument design (testing phase) or in phase 7 for the dissemination, impact and monitoring (stability phase).

The ability to create a protocol for practitioners to implement as a step by step guide in the testing phase allows researchers to test the implementation in real life educational settings from a very early stage in the research cycle. After the initial pilot, the protocol and instruments can be deployed across cohorts of schools over a sequential time period with the results aggregated in the form of a prospective cumulative meta-analysis. The ability as a researcher to see variations among the effect sizes of the individual trial cohorts and then the overall trends can then inform the design of larger scale randomised controlled trials on a national level for promising interventions. As the trials are school led, teachers and school leaders can opt-in to participate and create an evidence base that is constantly evolving and increasing in sample size. As highlighted by Siddiqui et al. (2015) the trials should lower the cost compared to traditional researcher led studies, promote innovation and result in less drop out. Furthermore, engaging practitioners at an earlier stage will allow practical implementation issues to be resolved before large scale studies are considered, reducing the risk of

attrition and school drop-out. The EFF recruitment and retention guidance provides researchers with advice on best practice on how to minimise this risk.

The second potential deployment of aggregated trials linked to a prospective cumulative meta-analysis is in the dissemination phase of the traditional research cycle. Usually when a large scale randomised controlled trial is completed, the results are disseminated to practitioners. In order to strengthen the external validity of the evidence, creating protocols to allow the research to be replicated when schools implement the intervention, they are able to use the same instruments and under the supervision of an independent researcher, ongoing data can be aggregated. Once sufficiency is established and a stable effect size is evident, a new prospective meta-analysis could be generated comparing educational interventions, rather than a control group. The advantage of using the aggregated trials linked to a PCM is the ability to observe and interpret the heterogeneity within each of the cohorts and in the PCM effect size for the intervention. As a researcher, it becomes possible to understand how the intervention is replaced and the larger the heterogeneity the greater the risk which is associated with scaling the intervention. Alternatively, the PCM could extend into different sample populations within schools or focus on systematically om different academic subjects.

10.3 Discussion on the why social sciences should adopt the use of protocols in evidence-based education

The global increase in the use of evidence in education is based on the premise that if programmes are selected on the basis of more robust evidence, using these tested interventions should increase the chance of positive outcomes if they are deployed in other schools and contexts. In other words, if we have robust evidence that an intervention has worked before, as a school leader looking to implement a similar intervention this would be a good bet that this would provide a positive outcome. As school budgets and resources are becoming more limited due to financial constraints, better informed decisions should help improve educational standards.

It becomes imperative that the evidence within education should be of the highest quality, accurate and free from bias to provide practitioners and policy makers an evidence base which they can use to inform their decisions. Randomised controlled trials are considered by the EEF and WWC as the ‘gold standard’ and the strength of a well implemented RCT lies in the ability to combine studies through systematic reviews and meta-analyses because of the strong internal validity of these studies. Consequently, as researchers we must ensure that the quality of the research produced is at the

highest possible standard. In order to provide better informed decisions, as a research community we need to address the current 'replication crisis' within educational research. Open Science Collaboration (2015) argue that "Innovation points out paths that are possible; replication points out paths that are likely; progress relies on both". If we are to generate an evidence base for practitioners, it is fundamental to look at ways to see if these results are reproducible, in order to provide the confidence that a particular intervention is likely to succeed.

It is perhaps more important to identify which features can be improved through developing research quality, and which are predictable consequences of particular designs. One possible solution might be to improve training for researchers into research design methodology when they start their research careers.

In light of the evidence suggesting improvements need to be made to provide the high quality, accurate and bias free evidence base for practitioners we need to consider how this can be achieved. As previously discussed, the EEF provides an example on how to conduct rigorous research to show what works in education. All studies are required to publish a protocol and statistical analysis plan before starting the research study. Independent evaluation teams then complete the evaluation and publish their analysis alongside their initial protocols and SAP's, enabling anyone to access these to independently evaluate the research. The transparency is a key strength for the research commissioned by the EEF and the results feed into a teaching and learning toolkit and more detailed guidance reports to enable dissemination to classroom practitioners and school leaders.

However, the processes in place to ensure rigour with the research commissioned by the EEF are not common practice for educational research. Yet examples exist for the use of protocols when topics span education and health, such as Chisholm et al., (2012) for an RCT for an educational school-based mental health intervention and Conner et al., (2013) RCT to reduce smoking initiation in adolescents. A further example of best practice involved a working example of the use of CONSORT to improve the reporting in RCTs in the Every Child Counts study by Torgerson et al., (2013). These example though are few and far between in educational research. Is this lack of transparency due to a wider knowledge gap in the academic research community outside of the research community focused on experimental designs?

` As a relatively inexperienced academic commencing the data collection for this thesis in the first year of the PhD, I was aware of the need to create a protocol and statistical analysis plan for the research. However, it was not until I fully understand the limitations of outcome reporting bias that I understood the importance of why these had to be published prior to the start of the research commencing. However, as a PhD student with limited financial funds to complete the research it was

not possible to fund the £226 + VAT per trial costs (ISRCTN, 2018) of registering these and as the data collection had commenced, the decision was made not to register retrospectively. Consequently, this limitation will be discussed in further detail in this current chapter.

However, the issue raises a wider concern in the financial constraints many researchers often encounter when conducting research, particularly early stage researchers who do not have access to funding. A potential solution to this would be requirement for all ethics boards in universities to insist that a protocol and statistical analysis plan should be submitted when ethics approval is sort. When the research is completed, publishers would be able to request these from the relevant ethics boards and make these available in the peer review process. If accepted for publication, these are then published alongside the article allowing full transparency in the research. This simple process will ensure a cost-effective approach for both the researchers and publishers as a short-term solution to tackle problem of outcome reporting bias. On the other hand, the solution prevents outcome reporting bias but the issue relating to publication bias is still a major risk to creating a more robust evidence base. The registration of all research will also allow the academic community to track incomplete studies, which is important to track as ethically time and funding goes into completing research.

The presence of publication bias is systematic review and meta-analyses is a widely accepted problem across a number of research disciplines (Banks, Kepes, & Banks, 2012; Dickersin, 1997; Dwan et al., 2008). In a review involving 645 studies from 12 reviews of reading, mathematics and science Cheung & Slavin (2015) found the overall effect sizes for published articles and unpublished reports were +0.30 and +0.16 respectively. The research evidence into possible reasons for publication bias and methods to minimise the problem are already discussed in Chapter 9 of this thesis. Yet, as an early stage researcher a simple solution using technology could present an opportunity for educational research to become fully transparent using blockchain technology to create a trials registry for all forms of educational research. Blockchain technology is also known as distributed ledger technology and creates a model whereby the ledger entries cannot be edited once these have been published (Tschorsch & Scheuermann, 2016). If a central blockchain repository is created for educational research, users are able to publish their protocols and statistical analysis plans for research that they are planning to complete. Due to the nature of the technology, if a researcher tried to amend the first entry this is prevented and they must upload the edited protocol as a separate entry and hence preventing publication bias. Combining the blockchain with a simple search repository would allow researchers to see particular areas of education for unpublished research. In addition, educational publishers would be able to cross reference submissions against initial research protocols and prevent duplication of unpublished research.

Importantly, the use of protocols should not be limited to educational research involving randomised controlled trials or quasi-experiments. In order for full transparency within the educational research evidence base, all research designs should record their research intentions in a central repository prior to research commencing. In clinical studies, the STROBE statement (Strengthening the Reporting of Observational Studies in Epidemiology) provides a framework for observational studies to improve reporting and transparency (von Elm, Altman, Egger et al., 2008).

In addition to the procedures outlined above, the creation of protocols without a standardised guidance will not necessarily improve reporting standards. In clinical trials lessons have been learned on the implementation of CONSORT in transforming the reporting and design of studies. In educational research, it is evident that we lack specific reporting guidelines in reporting educational research, regardless of research design. However, when reporting RCTs in education many researchers reference CONSORT as best practice and will publish protocols and SAPs with their research as best practice. Yet, these guidelines are predominately designed for clinical trials and as a research community we must develop a version for the social sciences.

In addition to the steps above to implement standardised protocols in educational research, the educational research journal editors should collectively provide guidance on minimum expectations to ensure published research is conducted to the highest quality, is free from bias and be accurate. The current model of peer review is fundamentally flawed due to the lack of pre-research registration (not just in trials), publication of protocols and analysis plans as it is impossible for external reviewers to assess the validity of the research without these.

10.4 Proposed theoretical model for using aggregated trials linked to a prospective cumulative meta-analysis in evidence-based education

It is now clear that it is imperative that the evidence within education should be of the highest quality, accurate and free from bias to enable practitioners and policy makers an evidence base which they can use to inform their decisions. Yet an argument should be that such a requirement is as important to fellow researchers relating directly to the first stage of the research cycle, as flawed and poorly informed literature will then translate to large scale studies and a waste of resource. Interventions that are designed based on unreliable research are unlikely to be effective, even if the design of the study is excellent and the intervention is implemented correctly. As previously explained, the EEF spends approximately £500,000 per trial (EEF, 2015) and the average effect sizes from an analysis of these high quality RCTs showed an average effect size of 0.06 standard deviations (Lortie Forgues & Inglis, 2019). The authors provide three possibly complimentary

possibilities on why this might be the case, as all studies included in the meta-analysis were methodologically sound. Firstly, it is possible that due to the lack of replication there is a possibility that the literature on which educational interventions are based are unreliable. As per the previous comment, a flawed piece of literature will translate into an ineffective trial result. Secondly, the translation of insights from the basic research evidence on which the trials are based are not translated into the successful implementation of the intervention. As explained in further detail in the limitations section of this chapter, a limitation with the primary research conducted in this thesis is the close monitoring and support of the researcher during the implementation of the intervention. The support ensured the intervention was delivered with a high fidelity, however if the trial had been at scale several implementation issues would present challenges to the fidelity of the study. Lortie-Forgues & Inglis (2019) suggest that a reason why many EEF and NCEE trials do not produce the expected effect sizes when a study is scaled as a large RCT is due to the lack the translation in the initial insights during the implementation in the larger study. As more schools are included at scale, the intervention is less likely to be implemented consistently across these. Slavin & Smith (2019) and Cronbach et al., (1980) refer to 'super realisation' as the fact that in small scale studies, researchers are able to monitor the quality of implementation and provide additional assistance creating unrealistic conditions. Finally, a further possibility is that background noise is underestimated in the research designs and that many of the trials are underpowered because of this.

If we consider the research cycle (Gorard, 2013) the seven phases progress from evidence synthesis, development of an idea or artefact, feasibility studies, prototyping and trialling, field studies and instrument design, rigorous testing and finally dissemination, impact and monitoring in phase 7. A similar research cycle is used by the EEF to allow applicants to provide evidence for the principles behind the intervention and of effectiveness in a peer reviewed grant application. A number of EEF funded studies have produced a promising effect sizes in efficacy trials but when these have progressed onto a large scale RCT the positive effect sizes have not translated. Recent examples include CatchUp®Numeracy (Hodgen et al. 2019), CatchUp®Literacy (Roy et al. 2019) and Grammar for Writing (Tracey et al. 2019). The reasons for his may relate to two potential key issues with the research cycle phases 3 to 5 covering the feasibility studies, prototyping and trialling and field studies (efficacy studies).

Firstly, in the phases 3 to 5 the purpose of these are to generate evidence for the principles behind the intervention and demonstrate a positive impact. Due to the design of feasibility studies and field trials, it is plausible that these will be small scale and tightly controlled in the implementation of the intervention. The schools involved are likely to receive additional support and

the intervention should be delivered with a high degree of fidelity. If a positive effect size is found and the design is robust, then the study is likely to be further funded and progressed into the rigorous testing phase of the research cycle. At this point, it is unlikely these studies will be replicated, and this is flaw with the current model. A positive effect size could have a large heterogeneity which is not able to be understood from a single feasibility or efficacy study, even if the research design is robust.

Secondly, as explained by Lortie-Forgues & Inglis (2019), the issue with scaling from a small tightly controlled study to a large scale RCT involving hundreds of schools will create issues replicating the translation of the insights from the smaller scale implementation. It is plausible that the intervention is not implemented consistently. A possible explanation for the inconsistency in schools could be the disconnect between teachers and understanding educational research (Lysenko et al., 2014; Levin, 2013). In educational trials, the implementation of the intervention will often rely on the teaching profession following the research guidance on how to implement the intervention. A lack of understanding or a failure to provide sufficient guidance will create an inconsistent approach across multiple schools.

10.4.1 Theoretical model for a research cycle linked to a PCM

A simplified version of the research cycle proposes to combine the phases 3 – 5 into a testing phase using clearly defined protocols for small scale aggregated trials linked to a PCM. Engaging practitioners earlier in cycle allows researchers to create a protocol for the study, instruments for testing and a step by step practitioner guide for implementing the intervention in schools. An initial cohort creates the pilot, to allow researchers and practitioners to modify materials based on practical implementation issues discovered in the first wave of the PCM. The second, third and future cohorts provide the opportunity to replicate the findings from cohort one, enabling researchers to understand if the effect size is stable and how heterogeneity is likely to impact if the study is scaled. If the heterogeneity is large across the PCM, then there is a risk that if scaled the intervention is likely to produce results that are uninformative. Furthermore, through the use of practitioner implementation protocols the risk of an inconsistent approach to the delivery of the intervention is reduced. As previously discussed, teachers often deliver interventions daily and providing a structured approach we increase the likelihood that teachers will buy into the need to follow a ‘recipe’ for a successful implementation. A useful analogy can be found from the popular TV show the Great British Bake-Off in their technical challenge, whereby a number of contestants

receive the same ingredients but very vague instructions on how to bake a specific bakery product. Rarely do the contestants produce the same texture or appearance, yet they all started with the same ingredients.

Through the development of a trial protocol and practitioner implementation guide we are scaffolding the knowledge learned from the initial promising pilot and upskilling teachers in how to conduct robust evaluations. Furthermore, through using cohorts of schools over a sequential time period we are able to replicate research. Makel & Plucker (2014) found that only 0.13% of articles in leading educational research journals are replicated studies, which increases the risk of flawed research being used as a basis for the foundation of an intervention. Figure 28 outlines a revised version of the research model proposed by Gorard (2013) to incorporate the use of PCM.

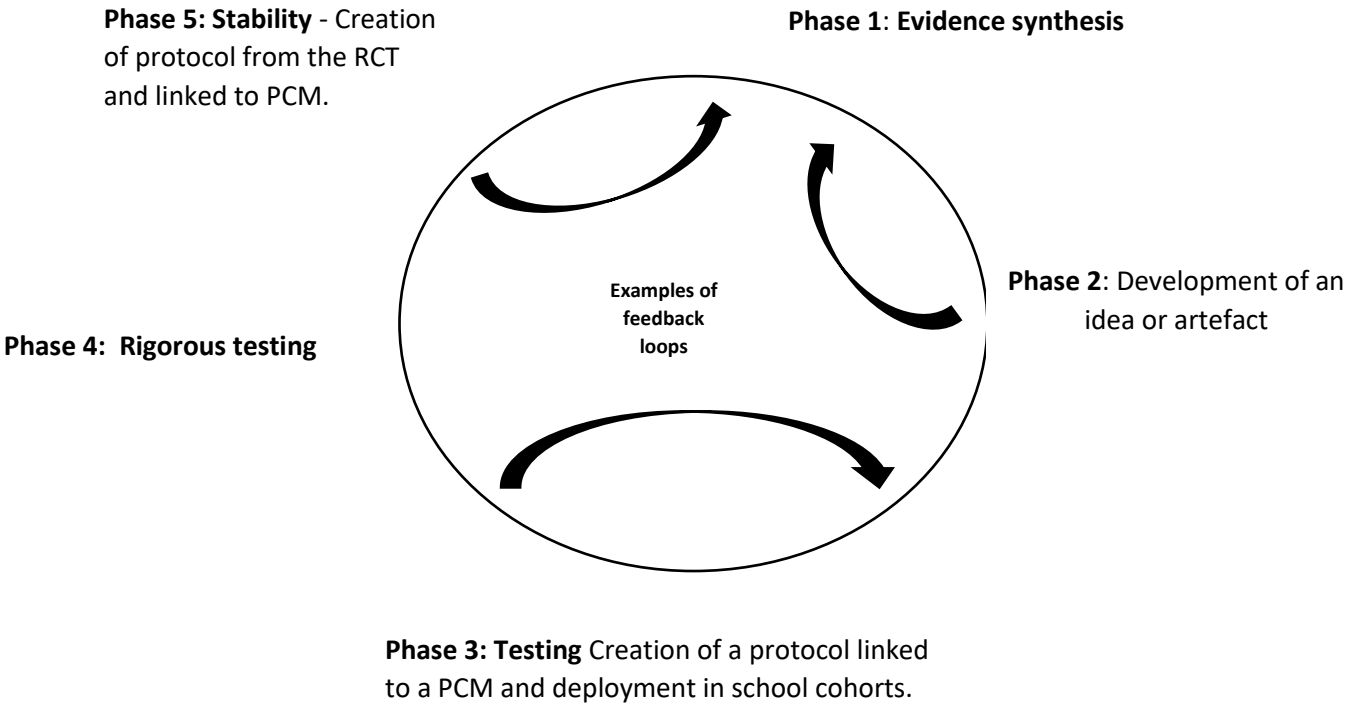


Figure 28: Adapted version of the research cycle created by Gorard (2013)

Through the implementation of a structured protocol and aggregated small-scale studies, organisations such as the EEF and NCEE are able to be better informed on the likelihood that an effect size will translate from the testing phase to the rigorous large scale RCTs. Ethically this is an important point, as large scale RCTs often require significant funding and resources, so the evidence base on which these are based must be as accurate and robust as possible. In addition to the use of a PCM in

the testing phase, the model proposes to replace the dissemination, impact and monitoring phase with a stability phase. The rationale for the change is linked to the current issues with dissemination and replication in educational research. If an intervention is tested through a large scale RCT, the findings are only generalisable to the population that has been tested and at the one point in time. For example, if a trial into the effectiveness of a mathematics catch-up scheme for data handling with children aged 11 – 13 is completed, can we apply this to algebra or a different age range of pupils? If protocols were created in dissemination of a traditional RCT trial and schools looking to adopt a peer tutoring strategy in their school, they could register to become part of a PCM cohort. The system creates a continuous flow of data until sufficiency is established. Direct replication can be tested using the same protocols for the RCT and instruments, allowing the stability of the effect size to be monitored. Furthermore, if schools were interested in testing the intervention on a different area of mathematics or with different target populations of learners, separate branches of the PCM will allow a greater understanding on which strategies work best and for whom.

The term sufficiency can potentially be a contentious point with relation to the use of a PCM, as who decides when we have sufficient evidence to confidently predict that a particular intervention will provide a positive effect size with a target population.

For practitioners, the ability to manage a small-scale intervention by following protocols allows the profession to develop their understanding of evidence and see how their data compares with the pool of evidence created for that specific PCM. An important point to reflect on at this point is the difference between the aggregated trials using clear protocols linked to a PCM and the research area of action research.

Kuhne et al. (1997) describe action research as an approach to problem posing and solving that progresses through four distinct processes: planning, acting, observing and reflecting. These create a cycle of research leading to another cycle of the four processes to address a problem. The knowledge gained by the teacher is used to inform their practice and the model allows revisions in each new cycle. The proposed model for protocols linked to aggregated trials and a PCM differ as the protocol cycle has to completed and cannot be broken or deviated. The data in the PCM model allows the practitioners to not only understand the impact in their own instance but also contribute to a wider evidence base. Critically practitioners can reflect on key questions such as ‘how does their trial compare with the pooled evidence?’ and ‘why has the effect size from the intervention differed from the expected?’

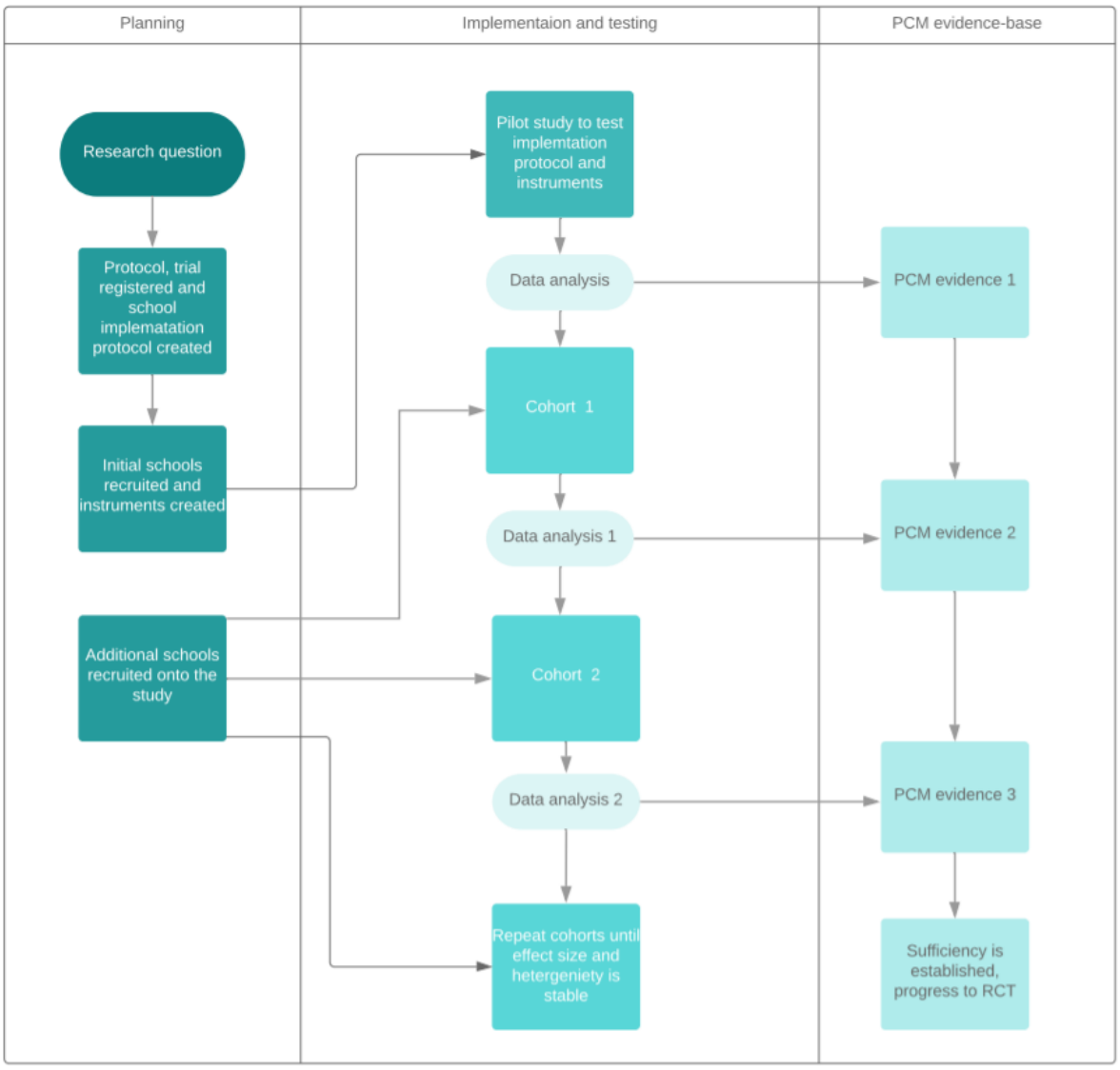


Figure 29: A model for the use of a standardised protocol for the aggregation small scale RCTs to create a PCM evidence base for a specific intervention in the testing phase of the research cycle.

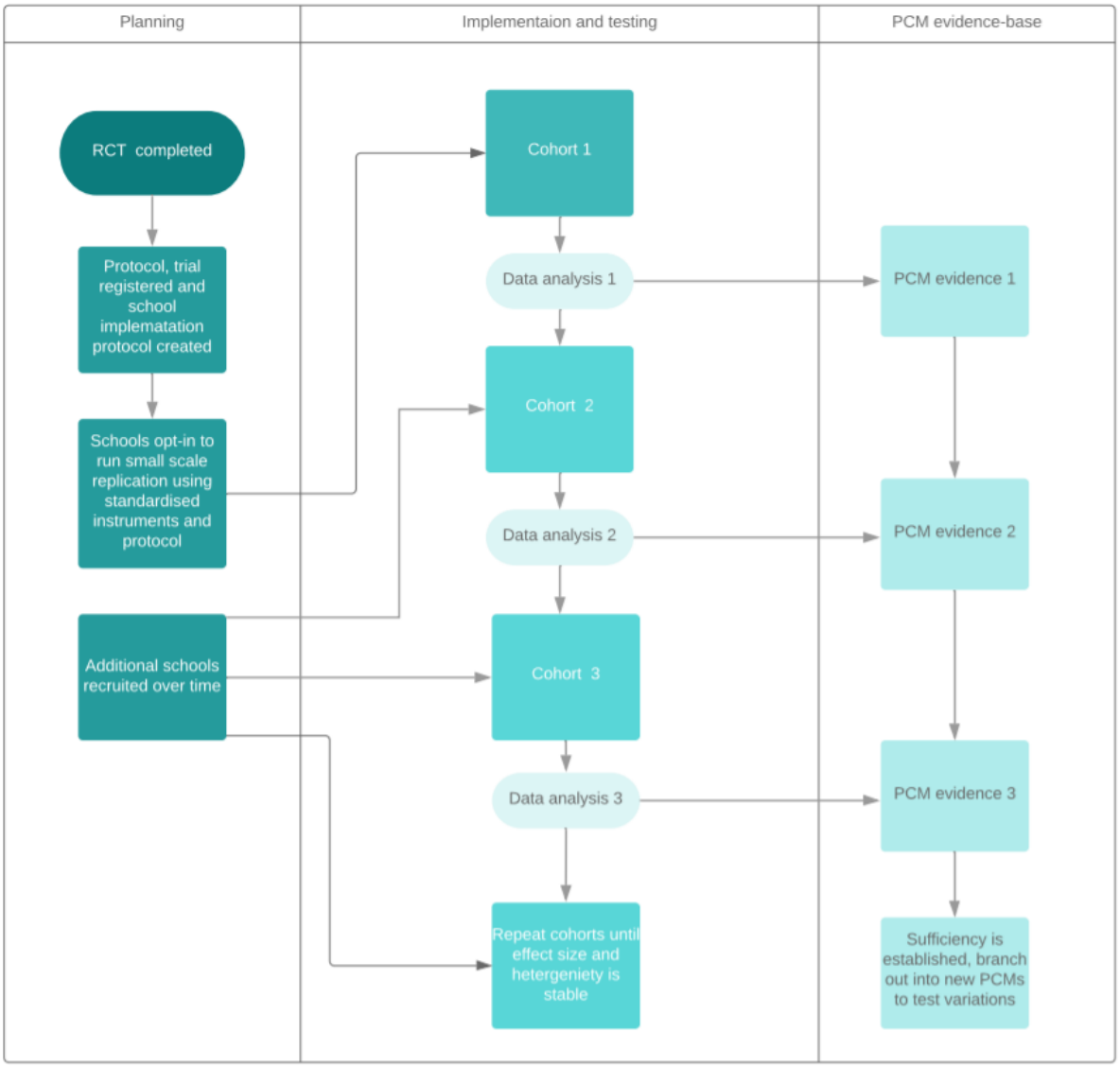


Figure 30: A model for the use of a standardised protocol for the aggregation small scale RCTs to create a PCM evidence base for a specific intervention in the stability phase of the research cycle.

The models presented for the testing and stability phase of the evidence generation demonstrate how replication is embedded into the process. Using standardised protocols for design and school-based implementation guides, the same instruments and similar target populations (such as Year 9 pupils currently working below expected progress) we are able to increase the likelihood that the implementation can be consistently replicated.

10.5 Limitations: Theoretical, Methodological and Implementation

10.5.1 Theoretical limitations

Firstly, it is important to acknowledge a limitation is the theoretical frameworks throughout the thesis in relation to the primary research studies. Both studies were conducted through the lens of a pragmatic framework, focusing on the design of the research, implementation and process evaluation for the aggregation of data to form the basis of a cumulative meta-analysis. Furthermore, the epistemological and ontological positions have not been covered and these are outside of the scope of the thesis.

At a theoretical level, peer tutoring is among many forms of peer learning techniques and Topping and Ehly (1998) provide a useful typology of these methods by classifying most peer assisted learning strategies as either:

1) Peer facilitation as observed in peer counselling;
2) Peer feedback such as peers monitoring and assessing each other and:
3) Peer tutoring such as same-age reciprocal peer tutoring, paired learning (same or cross-age tutoring) and one to one interactions involved in same-age class wide peer tutoring.

Consequently, the thesis does not explore either the traditional views such as Role Theory, Self-Determination Theory, Social Skills Theory or Piaget's Constructivists Theory or current and dominant theories relating to Vygotsky's Social Constructivist Theory and Social Independence Theories.

Both primary studies involved the use of online delivery for either peer to peer learning and small group teaching. If time permitted, a literature review on the online delivery model could provide a basis for developing a theoretical model for online teaching strategies, for whether this be peer or teacher led tuition.

The purpose of this thesis is to contribute to the knowledge within educational research through proposing a new methodology to provide a model to better understand the testing process for RCTs and the stability of research findings.

10.5.2 Methodological limitations

The primary studies were conducted in the first year of the current thesis as a follow on from the feasibility study conducted as part of the initial research for an MA in Research Methods. As the

platform for the online interventions had been made available, a pragmatic decision was made to move straight into a data collection period for the research due to the potential limited period of access to the technology. As the initial study involved three cohorts of schools over the academic year 2015 – 16, the research replicated the methodology from the previous research (Harrison, 2015) with the addition of a PCM and a sequential data collection throughout the year. The opportunity was presented to replicate the creation of a PCM for online small group teaching, extending the data collection for a further two terms in the academic year 2016 – 2017. Unfortunately, it was not possible to continue further than cohort 1 so all data had been collected within the expected period of one academic year.

The first methodological limitation is the lack of knowledge at the outset of the research for the need for trial registration. A trial protocol and SAP were created for both primary studies, using previous EEF studies and guidance as best practice. Yet, the reasons for the creation of these were not fully understood until the literature review on evidence-based education had been conducted. The In hindsight, delaying the primary data collection could possibly resulted in a better understanding of how outcome reporting bias (Bumble, 2011; Dwan et al., 2013) can be prevented within research practice through the registration of trials and the publication of protocols and SAPs. Even though a protocol and SAP had been produced, these were not published prior to the start of the data collection.

As the purpose of the two primary studies were to test the feasibility of the methodology for a PCM linked to small scale aggregated RCTs and protocols, the decision was made not to retrospectively register the trials. Firstly, the studies were small scale and not intended to generalise outside of the sample population. Secondly, the cost to register the trials post-data collection was probative due to a limited research budget. As the focus of the thesis involved the development of the methodology, lessons will be learned for future research and trial will be registered before the data collection period starts.

The pragmatic nature and limited finances for the research also provided a second methodological limitation in the use of non-standardised assessments for both the primary studies involved in this thesis. The importance of standardised assessments was evident in the research design for the previous research (Harrison, 2015). The independent InCAS assessments were previously used to reduce the potential impact for inflated effect sizes by using researcher made assessments compared to standardised assessments (Scammacca, 2007; Edmonds et al., 2009; Cheung & Slavin, 2015). However, the cost for access to these assessments for the current studies required alternatives to be used. In study one, previous SAT assessment questions were used by a primary teacher to help

create an assessment and a RASCH analysis was completed after the first cohort to check the reliability of the assessment. For study 2, an independent computer-based assessment programme called Alfiesoft was provided by the educational technology company that created the online platform. It is important to note that the questions selected for the assessment were selected by the technology company, so a potential source of bias could not be reduced to the pragmatic decision to use assessments that were not standardised.

Methodological, the primary research was not at the standard of a CONSORT trial, yet the research was conducted with the best intentions to conduct high quality research with limited resources. For example, a limitation within the methodology involved the randomisation process. As the same methodology was used from the previous study, randomisation was not conducted by a computer generation sequence software so the process was not recorded. Potential bias can occur if computer random assignment is not used.

The sample size within both primary studies is a limitation, as the research evidence demonstrates that studies with small sample sizes generally tend to have larger effect sizes than studies with larger sample sizes (Cheung & Slavin, 2015; Slavin, Lake & Groff, 2009). It is plausible that the effect sizes for both studies were affected by 'super realisation' as the researcher was able to monitor the quality of implementation and provide additional assistance creating unrealistic conditions at scale (Cronbach et al., 1980). Replicating the primary studies using a PCM without close researcher involvement will enable the potential impact of super realisation to be mitigated.

In both the primary studies, the statistical analysis plan created a limitation as the studies had originally planned to complete sub-group analyses of the differential impact of the interventions on pupil premium students. However it became clear that the sample size in both studies was not sufficient to complete this analysis.

Finally, the primary research studies could potentially be susceptible to the 'Hawthorne effect', whereby the research participants alter their behaviour as they were aware of being observed by the researcher. Due to the close involvement of the researcher in the implementation of the intervention throughout the data collection period, it is plausible that the participants behaviour was influenced providing a positive impact on the implementation of the intervention.

10.5.3 Implementation limitations

The process evaluation for the online peer tutoring and online small group teaching encountered similar issues during the implementation of the online interventions. The process evaluations in chapters 5 and 7 detail the implementation issues encountered in the research.

The main barriers to the implementation included: technical, logistical and financial constraints. Online educational interventions are reliant on a quality IT infrastructure within the settings of the delivery of the intervention, in both primary studies this involved school IT infrastructure. Overall the IT worked well for the duration of the data collection period, possibly due to a six-week planning period built into the trial design to allow school IT firewall settings to be configured and IT rooms to be scheduled for the delivery. In practical terms of delivering the intervention at scale, IT infrastructure is an important limitation as time is required to ensure the technology works before the intervention can be delivered. Consequently, the importance of understanding the IT requirements links to the second barrier, logistics. Educational environments such as schools are complex and are constrained by access to time for delivery, staffing requirements for supervision and access to IT facilities at a time when the intervention is delivered. In particular, the online peer tutoring required simultaneous access to computers at a specific time and day in two independent schools.

The implementation of the research is a strength rather than a weakness in this thesis, as a single researcher on a minimal research budget delivered two small scale RCTS with one study demonstrating the ability to aggregate data in a PCM. Using a protocol which outlined the implementation process, it was possible to mitigate many of the potential limitations that could impact on the quality of the research.

10.6 Recommendations

The next section of the chapter will discuss the key recommendations from the research to inform the design of future studies into online peer tutoring and small group learning, the lessons learned as a single researcher conducting an aggregated trial and recommendations for future research methodologies using cumulative meta-analyses. It also includes some reflections and speculation about how small-scale cumulative trials could be made practical in schools.

10.6.1 Recommendations for future research using online peer tutoring or small group teaching interventions

The process evaluations for both the primary studies found reoccurring themes in the barriers for delivering online interventions in school environments. Firstly, technical barriers are a fundamental component in the success of the intervention. It is important to assess the quality of the IT infrastructure before a school starts the onboarding process for the trial. The key areas to consider are the quality of the internet (bandwidth), firewall permissions and technical proficiency of a school IT team to clear these (particularly in primary schools) and the availability of suitable IT equipment such as computers. Secondly, logistical barriers can create issues in planning the delivery of interventions if these are delivered in school. It is important to discuss assess to computer rooms, availability of headsets if these are required and staffing requirements to supervise the intervention. Finally, financial barriers need to be discussed if the school is looking to continue with the intervention after the initial research is completed, particularly if third party software is used in the delivery of the intervention.

Further research is required to determine the effectiveness of online peer tuition across the transition boundary between primary and secondary schools, either as a continuation of the group size comparison or against a business as usual control. The intervention has proven the ability to use online technology to mediate an academic intervention for mathematics for data handling. Future research should explore how the pupils perceptions and confidence are impacted through using peer to peer interactions as an academic intervention. Focus groups and conservations with learners in all schools suggested a positive impact, however this cannot be generalised as the study was not designed to capture this evidence.

The Affordable Maths Tuition (Torgerson, 2016) evaluation recommended further research to explore online small group learning as this might be a more efficient and cost-effective method of tuition for schools. The current study adds to the limited research evidence for the impact of online small group teaching, even though the premature halt to the research after the first cohort limited the sample size. Consequently, further research is required to evaluate the effectiveness of small group teaching. In addition to a focus on mathematics, online small group teaching for academic subjects such as English, Science and MFL should be considered. The advantage of developing a PCM is that branches can evolve to test the effectiveness of the intervention, in his instance online small group teaching, in variations such as subject matter.

10.6.2 Recommendations for single researchers conducting an aggregated trial over a sequential time period

As a single researcher conducting primary research with a minimal budget and limited resources, it has been important to take a pragmatic approach to the design, implementation and process evaluation in each of the primary studies.

The most important lesson learned in managing a series of small-scale trials across numerous schools involved the constant communication with schools throughout the duration of the data collection period to coordinate the planning and delivery stage of each cohort, aligning this to fit the terms linked to the academic year. Establishing positive working relationships at each of the partner schools allowed the researcher to proactively resolve implementation issues to ensure that the intervention fidelity was high throughout the trial.

10.6.3 Recommendations for future research on the methodologies using cumulative meta-analyses

The next logical step in the development of prospective cumulative meta-analyses in education is the development of simulation studies to model the data for analysis. For example, previous trial data from a large-scale peer tutoring programme could be randomly sampled to form theoretical sample cohorts for the PCM. The analysis of the data would then create the individual cohort effect sizes and allow the distributions to be analysed for each wave. It is important to acknowledge the purpose of the study is purely theoretical and no generalisations relating the impact of the intervention could be inferred.

The evidence presented by Lortie-Forgues & Inglis (2019) shows that the current model of deciding on which promising studies to scale into larger RCTs at the EEF and NCEE evaluations may not be based on reliable evidence. Whether this is due to either flawed initial literature or hyper-realisation due to small scale studies, we need to rethink how we replicate studies to understand if the effect sizes can be reproduced and if heterogeneity is a potential risk if the intervention is scaled. A recommendation to the EEF and NCEE is to test using a protocol and small-scale aggregated trials with a PCM for an intervention in the testing phase of the research cycle. The evidence generated by the PCM should provide greater confidence that the effect size will translate into a larger RCT, before moving onto the stability phase to disseminate best practice and create a constant flow of data from school settings into the PCM.

This thesis is the start of the journey in exploring how we can engage practitioners in the generation of evidence through the use of protocols and small-scale RCTs in schools, developing a

common purpose within the profession to create an evidence base for what worked and for who. Future research should engage teachers in the co-creation of research questions which relate directly to their own practice and to that of others, providing teachers with the opportunity to select an intervention which they can they evaluate in their own school settings which is both applicable and appropriate for their pupils (Higgins, 2018). The key to engaging practitioners will be in removing or reducing the logistical and organisational barriers, allowing easy access to sign up to participate, perhaps through a website. The ability for the teacher to select an appropriate intervention, download the protocol and step by step implementation guide with resources and assessments could reduce teacher workload rather than increase it. The only difference would be the delivery schedule, as small-scale trials could use wait list designs or within subject randomisation so that all pupils do not receive the intervention at the same time. If such a website allowed the upload of anonymised data to the lead researcher, randomisation and data analysis would be independent of the teacher and could automatically accumulate, building up a picture of both the overall impact, but also the variability.

Crucially, as the resources, assessments and eligibility criteria are standardised in the initial pilot, these can then be replicated across numerous schools, increasing the sample size and providing multiple cohorts in each prospective cumulative meta-analysis. These rapid in-school evaluations would then supplement the evidence from synthesis and simulation studies to provide a better understanding of the potential of a cumulative approach to evidence generation.

10.7 Conclusions and personal reflections

In medicine, doctors are striving to constantly improve the outcomes for patients through finding new treatments, drugs and combinations of therapies to find what works and for who. When new treatments are designed or invented, they undergo a series of trials before the interventions even reach the general public. Imagine if a drug company designed a new drug, released it onto the market (unlicensed) for children to use and these children became ill or died as a result of the drug… this does not happen for a reason.

Unlike healthcare, all children are compelled to have education. Government policy, new teaching trends or technology interventions are unleashed on millions of children yet very few people stop and ask the questions, “are these effective?” and “what evidence do we have to decide if these interventions are effective?”.

In my past experience as a teacher, before starting an academic career in research, I admit I had very little understanding of evidence. I have numerous examples of interventions deployed in my

classrooms, either initiated at government level or through school improvement plans. A few notable examples include learning style questionnaires to design VAK (visual, auditory and kinaesthetic) tasks for individual lesson planning, brain gym activities, national teaching frameworks and booster materials, whole school deployment of interactive whiteboards and more recently class purchases of tablets. As a teacher I often trialled new strategies, such as using webinars for revision outside of school hours and flipping the classroom by creating online video tutorials. In a few instances, I used a pre and post assessment to measure the impact of the interventions in my classroom but evaluations of interventions in schools I worked in were rare.

In education, we seem to have good intentions and if we think something will work, we deploy it whole school for all children. Yet the notion of good intentions would not be used in medicine, so why is it used in education?

Fitz-Gibbon (2004) explains two examples where good intentions actually harmed children. Scared Straight was a programme based on the theory that if children showing signs of delinquency were taken to prisons to be told by inmates how horrible prison was, they would be deterred from a life of crime. Many subjective responses such as “it kept me away from crime” were given regarding the effectiveness of the programme, but through the use of a control group left untreated (randomised) it was found that the programme led to more serious crimes and increased recidivism in the treatment group compared to the control (McCord, 1978, 2001). A second example explains how the good intentions of providing counsellors to provide psychological debriefing immediately after a trauma and accidents. This seems a logical process to support patients’ wellbeing. However, two randomised trails concluded that this will do more harm (Mayou et al., 2000; Wessely et al., 1998).

In medicine, numerous examples exist where good intentions led to disastrous consequences. In the 1940’s and 50’s, premature babies were given pure oxygen (again, good intentions) yet during this time it was noted that there was an ‘epidemic’ of blindness among premature infants. It was also common for premature infants to be given prophylactic antibiotics, with an increase in brain damage and death recorded at this time. It was only through the use of randomised controlled trials that these consequences were identified and medical practice stopped delivering these interventions.

More recently, systematic reviews and meta-analyses have identified ineffective and harmful medical interventions, leading to doctors using evidence to inform prescribing of medications for patients. Yet, in education many seem to be happy to stay with the concept of ‘good intentions’ and even if evidence is used to find out the effectiveness of an intervention, often inferior ways to determine the effectiveness are used (See, 2018).

Fortunately, educational research is moving in the right direction towards an evidence-based practice approach to finding what has worked and what is likely to work if an intervention is used in school settings. The development of the EEF teaching and learning toolkit and the funding of large scale RCTs to test the effectiveness of interventions is a shining light in best practice for dissemination of previous research and for conducting RCTs using protocols, SAPs and independent evaluations. However, we still have issues when following in the clinical trials footsteps for implementing RCTs in educational settings. Morrison (2001) argues that education is too complex to reliably replicate any intervention, I agree that school environments are complex but I disagree that we cannot replicate a positive effect size from an intervention. If Morrison is correct, then we should stop all impact research as there would be no point in trying to inform practitioners or policy makers on what works. As with medicine, the impact of a lack of or poor-quality education can have profound consequences on the life chances of children. It is clear from the literature (Coe et al., 2000; Gorard, 2013; Higgins, 2017) that evidence-based education does involve the assumption that when an intervention or change is implemented, a causal mechanism is inferred, and it is crucial to understand the counterfactual or comparison being made. Within the education-evidence base we should find a wealth of evidence using a wide range of research designs, from ethnographic studies to rigorous evaluations using RCTs. We should expect that all researchers have the common aim to create evidence of the highest quality, be accurate and free from bias to provide practitioners and policy makers an evidence base which they can use to inform their decisions.

As a relatively inexperienced academic, it is clear that without researchers registering their research and explaining using protocols how they intend to conduct their research, we have no transparency within the system. Whether a researcher is conducting a case study on the impact of peer led sex education or planning a quasi-experiment to investigate new maths intervention, the research has a causal mechanism so the research questions, methodology and instruments should be published prior to the start of the data collection. Without transparency, the risk increases of the likelihood that the research may be flawed through potential sources of bias.

Furthermore, even though the EEF and similar evidence-based organisations are leading the way in the social sciences for conducting large scale RCTs, further improvements can still be made. This thesis has proposed and tested the development of using protocols with small scale RCTs and aggregating the data into a PCM. The model proposes simplifying the current research cycles as proposed by Gorard (2013) to create a testing phase to allow small scale replication in the educational setting, such as researchers using practitioner involvement to develop implementation protocols for schools. The replication of the intervention across small cohorts allows researchers to test the reproducibility of the effect size and observe the heterogeneity within the PCM. If the effect

size and heterogeneity is stable, then the intervention is more likely to scale and translate to a positive effect size in a rigorous RCT. Furthermore, if the researchers conducting the RCT create a protocol for continued school testing, teachers will be able to recreate small scale evaluations to allow the stability of the intervention to become understood. It is crucial for researchers to involve practitioners in all aspects of the research cycle if we are to transform the profession so that evidence becomes the foundation on which decisions are made.

In conclusion, we would not allow large scale pharmaceutical companies to release untested drugs in society, so we must understand that in education it is equally as important to create an evidence-base which informs practice.

References

Adey, P., Hewitt, G., Hewitt, J., Landau, N. (2004). *The Professional Development of Teachers: Practice and Theory.* (London (ed.)). Kluwer Academic Publishers.

Ainsworth, H., Hewitt, C. E., Higgins, S., Wiggins, A., Torgerson, D. J., & Torgerson, C. J. (2014). Sources of bias in outcome assessment in randomised controlled trials: a case study. *Educational Research and Evaluation*, *3611*(January 2015), 1–12. https://doi.org/10.1080/13803611.2014.985316

Altman, D.G., Dore, C. J. (1990). Randomisation and baseline comparisons in clinical trials. *Lancet*, *335*, 149–153.

Andrew, E. (1994). A Proposal for Structured Reporting of Randomized Controlled Trials. *JAMA: The Journal of the American Medical Association*, *272*(24), 1926. https://doi.org/10.1001/jama.1994.03520240054041

Andrews, K., Stewart, J. (1979). Stroke recovery: he can but does he? *Rheumatology and Rehabilation*, *18*, 43–48.

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication Bias: The Antagonist of Meta-Analytic Reviews and Effective Policymaking. *Educational Evaluation and Policy Analysis*, *34*(3), 259–277. https://doi.org/10.3102/0162373712446144

Bannister, J. (2011). Building safer communities: Knowledge mobilisation and community safety in Scotland. *Crime Prevention and Community Safety*, *13*(4), 232.

Bassey, M., Constable, H. (1997). *Higher Education Research in Education, 1992-1996: Fields of Enquiry Reported the HEFCs RAE*.

Bassey, M. (1996, November). We are specialists at pursuing truth. *Times Educational Supplement*.

Becker, B. J. (2005). The failsafe N or file-drawer number. In *In H.R Rothstein, A.J. Sutton, & M. Borenstein (Eds). Publication bias in meta-analysis: prevention, assessment and adjustments* (pp. 111–126). UK: Wiley.

Begg, C., Cho, M., Eastwood, S., Els, D., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K. F., Simel, D., & Stroup, D. F. (1996). *Improving the quality of Randomized Controlled Trials: The CONSORT Statement*. *8*(D), 637–639.

Biesta, G. (2007). Why "what works" won't work: evidence-based practice and the democratic deficit in educational research. *Educational Theory*, *57*(1), 1–22. https://doi.org/10.1111/j.1741-5446.2006.00241.x

Blümle, A., Meerpohl, J. J., Rucker, G., Antes, G., Schumacher, M., & von Elm, E. (2011). Reporting of eligibility criteria of randomised trials: cohort study comparing trial protocols with subsequent articles. *Bmj*, *342*, d1828. https://doi.org/10.1136/bmj.d1828

Bowen, D. J., & Neill, J. T. (2013). A meta-analysis of adventure therapy outcomes and moderators [Data file]. Unpublished dataset, University of Canberra, Canberra, Australia. *The Open Psychology Journal*, *6*, 28–53.

Bradford-Hill, A. (1966). 'The environment and diseasse: Association or causation? *Proceedings of Royal Society of Medicine*, 58:285.

Branon, R., & Essex, C. (2001). Synchronous and asynchronous communication tools in distance education. *TechTrends*, *45*(1), 36. http://dx.doi.org/10.1007/BF02763377

Braun, V., & Clarke, V. (2006). *Using thematic analysis in psychology*. *3*, 77–101. https://doi.org/10.1191/1478088706qp063oa

Campbell, M., Fitzpatrick, R., Haines, a, Kin outh, A., Sandercock, P., Spiegelhalter, D., & Tyrer, P. (2000). Framework for design and evaluation of complex interventions to improve health. *Bmj*, *321*(7262), 694–696. https://doi.org/10.1136/bmj.321.7262.694

Cartwright, N. (2013). Knowing what we are talking about: Why evidence doesn't always travel. *Evidence and Policy*, *9*(1), 97–112. https://doi.org/10.1332/174426413X662581

Chan, A.-W., Hróbjartsson, A., Jørgensen, K. J., Gøtzsche, P. C., & Altman, D. G. (2008). Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ (Clinical Research Ed.)*, *337*, a2299. https://doi.org/10.1136/bmj.a2299

Chan, A.-W., Tetzlaff, J. M., Gøtzsche, P. C., Altman, D. G., Mann, H., Berlin, J. a, Dickersin, K., Hróbjartsson, A., Schulz, K. F., Parulekar, W. R., Krleza-Jeric, K., Laupacis, A., & Moher, D. (2013). SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ (Clinical Research Ed.)*, *346*, e7586. https://doi.org/10.1136/bmj.e7586

Chan, A.W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials. *The Journal of the American Medical Association*, *291*(20), 2457–2465.

Chan, An Wen, Tetzlaff, J. M., Altman, D. G., Laupacis, A., Gøtzsche, P. C., Krleža-Jerić, K., Hróbjartsson, A., Mann, H., Dickersin, K., Berlin, J. a., Doré, C. J., Parulekar, W. R., Summerskill, W. S. M., Groves, T., Schulz, K. F., Sox, H. C., Rockhold, F. W., Rennie, D., & Moher, D. (2013). SPIRIT 2013 statement: Defining standard protocol items for clinical trials. *Annals of Internal Medicine*, *158*(3), 200–207. https://doi.org/10.7507/1672-2531.20130256

Chappell, L., Alfirevic, Z., Chien, P., Jarvis, S., Thornton, J. G. (2005). A comparison of the published version of randomized controlled trials in a specialist clinical journal with the original trial protocols. *5th International Congresson Peer Review and Biomedical Publications*. http://www.peerreviewcongress.org/abstracts.html#randomized

Chedzoy, S.M., & Burden, R. L. (2005). Assessing student attitudes to primary–secondary school transfer. *Research in Education*, *74*, 22–35.

Cheung, A. C. ., & Slavin, R. . (2016). How Methodological Features Affect Effect Sizes in Education. *Educational Researcher*, *45*(5), 283–292.

Chinn, C.A., Brewer, W. F. (1993). The Role of Anomolous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction. *Review of Educational Research*, *63*(1), 1–49.

Chisholm, K. E., Patterson, P., Torgerson, C., Turner, E., & Birchwood, M. (2012). A randomised controlled feasibility trial for an educational school-based mental health intervention: Study protocol. *BMC Psychiatry*, *12*(August), 0–7. https://doi.org/10.1186/1471-244X-12-23

Ciesielska, M., Bostrum, K.W., Ohlander, M. (2018). Observation methods. In *Qualitative Methodologies in Organisational Studies.* (pp. 33–52).

Clapham, K., Manning, C., Williams, K., O'Brien, G., & Sutherland, M. (2017). Using a logic model to evaluate the Kids Together early education inclusion program for children with disabilities and additional needs. *Evaluation and Program Planning*, *61*, 96–105. https://doi.org/10.1016/j.evalprogplan.2016.12.004

Cochrane, A. L. (1972). *Effectiveness and efficiency: Random reflections on health services.*

Coe, R., Fitz-Gibbon, C., & Tymms, P. (2000). Promoting Evidence -Based Education: The Role of Practitioners. *Round Table Presented at the British Educational Research Association Annual Conference*, *4*(March), 1–10.

Cohen, P. A., Kulik, J. A., & Kulik, C. L. C. (1982). Educational Outcomes of Tutoring: A Meta-analysis of Findings. *American Educational Research Journal*, *19*(2), 237–248. https://doi.org/10.3102/00028312019002237

Cohen, P. a, Kulik, J. a., & Kulik, C.-L. C. (1982). Educational Outcomes of Tutoring: A Meta-analysis of Findings. *American Educational Research Journal*, *19*(2), 237–248. https://doi.org/10.3102/00028312019002237

Cole, H., & Griffiths, M. D. (2007). Social interactions in massively multiplayer online role-playing gamers. *Cyberpsychology & Behavior : The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *10*(4), 575–583. https://doi.org/10.1089/cpb.2007.9988

Collaboration, Campbell. (2019). *Campbell Collaboration*. www.campbellcollaboration.org

Collaboration, Cochrane. (2016). *Cochrane Methods*. http://methods.cochrane.org/pma/what-pma-0

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Cook, S.B., Scruggs, T.E., Mastropieri, M.A., & Castro, G. C. (1985). Handicapped students as tutors. *Journal of Special Education*, *19*, 483–492. https://doi.org/10.1177/002246698501900410

Cook, D. A. (2014). How much evidence does it take? A cumulative meta-analysis of outcomes of simulation-based education. *Medical Education*, *48*(8), 750–760. https://doi.org/10.1111/medu.12473

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Rand McNally.

Corbin, J., Morse, J. M. (2003). The unstructured interactive interview: Issues with reciprocity and risks when dealing with sensitive topics. *Qualitative Inquiry*, 335–353.

Cordingley, P. (2008). Research and evidence-informed practice: Focusing on practice and practitioners. *Cambridge Journal of Education*, 38(1), 37-52.

Cousins, J. B., & Leithwood, K. A. (1993). Current Empirical Research on Evaluation Utilization. *Review of Educational Research*, *56*(3), 331–364.

Cousins, J. B., & Walker, C. (2000). Predictors of educators' valuing of systematic inquiry in schools. *Canadian Journal of Program Evaluation*, *Special Ed*, 25–52. http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ610733

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.O., Hornik, R.C., Phillips, D. C. (1980). *Toward reform of program evaluation: Aims, methods, and institutonal arrangements*. Jossey-Bass.

Davies, P. (1999). What is Evidence-based Education? *British Journal of Educational Studies*, *1005*(December). https://doi.org/10.1111/1467-8527.00106

de Boer, H., Donker, a. S., & van der Werf, M. P. C. (2014). Effects of the Attributes of Educational Interventions on Students' Academic Performance: A Meta-Analysis. *Review of Educational Research*, 0034654314540006-. https://doi.org/10.3102/0034654314540006

Deaton, A., & Cartwright, N. (2016). Understanding and Misunderstanding Randomized Controlled Trials. *NBER Working Paper*, *44*(22595), 1–69. https://doi.org/10.3386/w22595

Dekhinet, R.,Topping, K.J., Duran, D., & Blanch, S. (2008). Let me learn with my peers online: Foreign

language learning through reciprocal peer tutor. *Innovate*. http://www.researchgate.net/profile/Silvia_Blanch/publication/234775503_Let_Me_Learn_with_My_Peers_Online%21_Foreign_Language_Learning_through_Reciprocal_Peer_Tutoring/links/00b7d529c3808413f6000000.pdf?origin=publication_detail

Denscombe, M. (2003). *The Good Research Guide: For Small-Scale Social Research Projects.* (Open Unive).

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Contr Clin Trials*, *7*, 177–188.

Dickersin, K. (1997). How important is publication bias? A synthesis of available data. *AIDS Education and Prevention*, *9*(January), 15–21.

Dumville, J.C., Torgerson, D.J., Hewitt, C. E. (2006). Reporting attrition in randomised controlled trials. *BMJ*, *322*, 969–971.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P. J., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J. P. A., Simes, J., & Williamson, P. R. (2008). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. *PLoS ONE*, *3*(8), e3081. https://doi.org/10.1371/journal.pone.0003081

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, a., Tackett, K. K., & Schnakenberg, J. W. (2009). A Synthesis of Reading Interventions and Effects on Reading Comprehension Outcomes for Older Struggling Readers. *Review of Educational Research*, *79*(1), 262–300. https://doi.org/10.3102/0034654308325998

Education Endowment Foundation. (2015). *EEF*. https://educationendowmentfoundation.org.uk/about/

EEF. (2016). *Education Endowment Foundation*. Evaulation Resources. https://educationendowmentfoundation.org.uk/evaluation/resources-centre/writing-a-protocol/

EEF. (2017). *Research Schools*. https://educationendowmentfoundation.org.uk/our-work/research-schools/

Elbaum, B., Vaughn, S., Hughes, M.T., Moody, S. W. (2000). Education as socialization and as individualization. *Journal of Educational Psychology*, *92*(4), 114–126. https://doi.org/10.1037//00224

Falloon, G. (2014). What's going on behind the screens? Researching young students' learning pathways using iPads. *Journal of Computer Assisted Learning*, *30*(4), 318–336. https://doi.org/10.1111/jcal.12044

Fitz-Gibbon, C. (2004). Editorial : The need for randomized trials in social. *Journal of the Royal Statistical Society*, *167*(1), 1–4.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Freemantle, N., & Watt, I. (1994). Dissemination: implementing the findings of research. *Health Libraries Review*, *11*(2), 133–137. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med3&NEWS=N&AN=10136651

Freiman, J.A., Chalmers, T.C., Smith, H Jr, Kuebler, R. . (1978). The importance of beta, the type II error, and sample size in the design and interpretation of randomised control trial: survey of 71'negative' trials. *N Engl J Med*, *299*, 690–694.

Ginsburg-Block, M. D., Rohrbeck, C., & Fantuzzo, J. W. (2006). A meta-analytic review of social, self-concept, and behavioral outcomes of peer-assisted learning. *Journal of Educational Psychology*, *98*(4), 732–749. https://doi.org/10.1037/0022-0663.98.4.732

Ginsburg, A., & Smith, M. S. (2016). Do Randomized Controlled Trials Meet the " Gold Standard "? A Study of the Usefulness of RCTs in the What Works Clearinghouse. *American Enterprise Institute*, *March*.

GISSI. (1986). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarctions. *Lancet*, *1*, 397–402.

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. *Midwest Research to Practice Conference in Adult, Continuing, and Community Education*, *1992*, 82–88. https://doi.org/10.1109/PROC.1975.9792

Goldarce, B., Drysdale, H. Powell-Smith, A., Dale, A., Milosevic, I., Slade, E., Marston., C., Heneghan, C. (2016). *The COMPare Trials Project*. www.COMPare-trials.org

Gorard, S., Siddiqui, N., See, B. H. (2015). Fresh Start. *Education Endowment Foundation.* https://educationendowmentfoundation.org.uk/uploads/pdf/Fresh_Start_(Final).pdf

Gorard, S., Siddiqui, N. and See, B. H. (2016). (2016). An evaluation of Fresh Start as a catch-up intervention: a trial conducted by teachers. *Educational Studies*, *42*(1), 98–113.

Gorard, S. & Taylor, C. (2004). *Combining Methods in Educational and Social Research*. London: Open University Press.

Gorard, S. (2013). *Research Design: Creating Robust Approaches for the Social Sciences*. Sage Publications.

Gough, D. (2013). Knowledge mobilisation in education in England. In *The Impact of Research in Education* (pp. 65–84). The Policy Press.

Graham, C., & Hill, M. (2003). *Negotiating the transition to secondary school.* (Spotlight). The SCRE Centre, University of Glasgow.

Gray, J., Goldstein, H., William, K. (1997, February). Educational Research and Evidence-based Practice: the debate continues. *Research Intelligence*, 18–20.

Gray, J. (1996). Track record of peer review: a reply to some remarks by David Hargreaves. *Research Intelligence*, Number 57, 5-6.

Greenburg, M.T., Domitrovich, C.E., Graczyk, P.A., Zins, J. E. (2005). *The study of implementation in school-based preventive interventions: Theory, practice and reserach.* (Issue March 2016).

Griggs, J., Speight, S., & Farias, J. C. (2016). *Ashford Teaching Alliance Research Champion*.

Grissom, R.J., & Kim, J. J. (2005). *Effect Sizes for Research: A Broad Practical Approach.* Mahwah, NJ: Lawerence Erlbaum.

Habler, B., Major, L., Hennessy, S. (2016). Tablet use in schools: a critical review of the evidence for learning outcomes. *Journal of Computer Assisted Learning*, *32*(2), 139–156.

Hammersley, M. (1997). Educational Research and Teaching: a response to David Hargreaves' TTA lecture. *British Educational Research Journal*, *23*(2), 141–161. https://doi.org/10.2307/1501807

Hargreaves, D. H. (1996). Teaching as a research-based profession: Possibilities and prospects. *The Teacher Training Agency Annual Lecture*, 1–12. https://doi.org/10.1017/CBO9781107415324.004

Hargreaves, David H. (1997). In Defence of Research for Evidence-based Teaching: a rejoinder to

Martyn Hammersley. *British Educational Research Journal*, *23*(4), 405–419. https://doi.org/10.1080/0141192970230402

Harrison, W. D. (2015). *A feasibility study: Online cross-age peer tuition across the transition boundaries of primary to secondary schools*. Durham University.

Have, M., Nielsen, J. H., Gejl, A. K., Thomsen Ernst, M., Fredens, K., Støckel, J. T., Wedderkopp, N., Domazet, S. L., Gudex, C., Grøntved, A., & Kristensen, P. L. (2016). Rationale and design of a randomized controlled trial examining the effect of classroom-based physical activity on math achievement. *BMC Public Health*, *16*(1), 304. https://doi.org/10.1186/s12889-016-2971-7

Hawkey, C. J. (2001). Journals should see original protocols for clinical trials. *BJM: British Medical Journal*, *323*(7324), 1309.

Hedges, H. (2012). Teachers' funds of knowledge: A challenge to evidence-based practice. *Teachers and Teaching*, *18*(1), 7-24.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods*, *1*(2), 154–169. https://doi.org/10.1037/1082-989X.1.2.154

Hemsley-Brown, J., & Sharp, C. (2003). The use of research to improve professional practice: a systematic review of the literature. *Oxford Review of Education*, *29*(4), 449–470. https://doi.org/10.1080/03054980701476311

Hemsley-Brown, J. V., & Sharp, C. (2004). The use of research to improve professional practice : a systematic review of the literature. *Oxford Review of Education*, *29*(4), 449–470.

Hextall, I., Mahony, P., Sidgewick, S. (1998). Chinese whispers and other party games. *Research Intelligence*, 19–20.

Higgins, S., Katsipataki, M., Coleman, R., Henderson, P., Major, L.E., & Coe, R. (2015). (2015). *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit.* London: Education Endowment Foundation.

Higgins, J. P. T., Whitehead, A., & Simmonds, M. (2011). Sequential methods for random-effects meta-analysis. *Statistics in Medicine*, *30*(9), 903–921. https://doi.org/10.1002/sim.4088

Higgins, S. (2017). Room in the Toolbox? The place of randomised controlled trials in educational research. In I. Childs, A., Menter (Ed.), *Mobilising Teacher Researchers: Challenging Educational Inequality* (pp. 97–1122). Routledge.

Higgins, S., Xiao, Z., & Katsipataki, M. (2012). *The Impact of Digital Technology on Learning : A Summary for the Education Endowment Foundation Full Report*. *November*, 1–52.

Hodkinson, P. (1998). Naivete and bias in educational research: the Tooley Report. *Research Intelligence*, 19–20.

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D. G., Barbour, V., Macdonald, H., Johnston, M., Lamb, S. E., Dixon-Woods, M., McCulloch, P., Wyatt, J. C., Chan, A.-W., & Michie, S. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ (Clinical Research Ed.)*, *348*(mar07_3), g1687. https://doi.org/10.1136/bmj.g1687

Horton, R. (2006). Trial registers: protecting patients, advancing trust. *Lancet*, *367*(9523), 1633–1635. https://doi.org/10.1038/news060313

Hrastinski, S. (2006). Introducing an informal synchronous medium in a distance learning course: How is participation affected? *Internet and Higher Education*, *9*, 117–131.

Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Rasmussen, J. V., Hilden, J., Boutron, I., Ravaud, P., & Brorson, S. (2014). Observer bias in randomized clinical trials with time-to-event outcomes: Systematic review of trials with both blinded and non-blinded outcome assessors. *International Journal of Epidemiology*, *43*(3), 937–948. https://doi.org/10.1093/ije/dyt270

Hultman, G., & Horberg. C.R. (1998). Knowledge Competition and Personal Ambition: A Theoretical Framework for Knowledge Utilization and Action in Context. *Science Communication*, *19*, 328–348.

Institute of Educational Sciences. (2015). *What Works Clearinghouse*. http://ies.ed.gov/ncee/wwc/aboutus.aspx

Ioannidis, J. P., Cappelleri, J. C., & Lau, J. (1998). Issues in comparisons between meta-analyses and large trials. *Jama*, *279*(14), 1089–1093. https://doi.org/10.1097/00132586-199906000-00063

ISIS2. (1988). Randomised trial of intravenous streptokinase, oral asprin, both, or neither among 17,187 cases of suspected cute myocarial infarctions. *Lancet*, *12*(6), 3A-13A.

Jones, R. H., Garralda, A., Li, D. C. S., & Lock, G. (2006). Interactional dynamics in on-line and face-to-face peer-tutoring sessions for second language writers. *Journal of Second Language Writing*, *15*(1), 1–23. https://doi.org/10.1016/j.jslw.2005.12.001

Jun, S. W., Ramirez, G., & Cumming, A. (2010). Tutoring Adolescents in Literacy: A Meta-Analysis. *McGill Journal of Education*, *45*(2), 219–238. http://libproxy.uwyo.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ908983&site=ehost-live&scope=cite

Kasim, A., Xiao, Z., Higgins, S., Troyer, E. D. (2017). *eefAnalytics: Analysing Education Trials* (1.0.6).

Kasim, A.D., ZhiMin, X., & Higgins, S. (2015). *Overview of analysis of EEF trials.*

Kicinski, M. (2013). Publication bias in recent meta-analyses. *PLoS ONE*, *8*(11), 1–10. https://doi.org/10.1371/journal.pone.0081823

Kivunja, C. (2015). Innovative Methodologies for 21st Century Learning, Teaching and Assessment: A Convenience Sampling Investigation into the Use of Social Media Technologies in Higher Education. *International Journal of Higher Education*, *4*(2), 1–26. https://doi.org/10.5430/ijhe.v4n2p1

Kowert, R., & Oldmeadow, J. a. (2013). (A)Social reputation: Exploring the relationship between online video game involvement and social competence. *Computers in Human Behavior*, *29*(4), 1872–1878. https://doi.org/10.1016/j.chb.2013.03.003

Kraemer, H.C., Blasey, C. (2016). *How Many Subjects? Statistical Power Analysis in Research*. Sage Publications.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PloS One*, *9*(9), e105825. https://doi.org/10.1371/journal.pone.0105825

Kuyath, S. (2008). The Social Presence Of Instant Messaging: Effects On Student Satisfaction, Perceived Learning, And Performance In Distance Education. *Ph.D. Thesis University of North Carolina at Charlotte*.

Kyaergard, L.L., Villumsen, J., & Gluud, C. (2001). Reported methodological quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of Internal Medicine*, *135*, 982–989.

Landry, R., Amara, N., Lamari, M. (2001). Climbing the ladder of research utilization: Evidence from social science research. *Science Communication*, *22*, 396–422.

Lau, J., Antman, E.M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., and Chalmers, T. C. (1992). The New England Journal of Medicine Downloaded from www.nejm.org at MOUNT SINAI SCHOOL OF MEDICINE on November 1, 2010. For personal use only. No other uses without permission. Copyright © 1992 Massachusetts Medical Society. All rights reserved. *The New England Journal of Medicine*, *327*(4), 248–254.

Lau, J., Schmid, C. H., & Chalmers, T. C. (1995). Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *Journal of Clinical Epidemiology*, *48*(1), 45–57. https://doi.org/10.1016/0895-4356(94)00106-Z

Leung, K.C. (2015). Preliminary empirical model of crucial determinants of best practice for peer tutoring on academic achievement. *Journal of Educational Psychology*, *107*(2), 558–579. https://doi.org/10.1037/a0037698

Leung, Kim Chau. (2015). *Preliminary Empirical Model of Crucial Determinants of Best Practice for Peer Tutoring on Academic Achievement*. *107*(2), 558–579.

Levin, B. (2013). To know is not enough: research knowledge and its use. *Review of Education*, *1*(1), 2–31. https://doi.org/10.1002/rev3.3001

Levin, Benjamin, Cooper, A., Arjomand, S., & Thompson, K. (2011). Can simple interventions increase research use in secondary schools? *Canadian Journal of Educational Administration and Policy*, *December*(126), 1–29. http://www.umanitoba.ca/publications/cjeap/archives.html %5Cnhttp://www.umanitoba.ca/publications/cjeap/pdf_files/levin-etal.pdf

Li, Q., & Ma, X. (2014). A Meta-analysis of the Effects of Computer Technology on School Students ' Mathematics Learning Author (s): Qing Li and Xin Ma Published by : Springer Stable URL : http://www.jstor.org/stable/23364142 . Your use of the JSTOR archive indicates your accep. *Educational Psychology Review*, *22*(3), 215–243. https://doi.org/10.1007/sl0648-010-9125-8

Liao, Y. K. C. (1999). Effects of hypermedia on students' achievement: a meta-analysis. *Journal of Educational Multimedia and Hypermedia*, *8*(3), 255–277.

Lin, W. C., & Yang, S. C. (2013). Exploring the roles of Google.doc and peer e-tutors in English writing. *English Teaching*, *12*(1), 79–90.

Lipsey, M., & Wilson, D. (1993). The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation From Meta-Analysis. *American Psychologist*, *48*(12), 1181–1209. https://doi.org/10.1037//0003-066X.48.12.1181

Littell, J. H., & White, H. (2018). The Campbell Collaboration: Providing Better Evidence for a Better World. *Research on Social Work Practice*, *28*(1), 6–12. https://doi.org/10.1177/1049731517703748

Lloyd, C., Edovald, T., Kiss, Z., Morris, S., Skipp, A., & Ahmed, H. (2015). Paired Reading. In *Education Endowment Foundation.* https://educationendowmentfoundation.org.uk/uploads/pdf/Paired_Reading.pdf

Lloyd, C., Edovald, T., Kiss, Z., Morris, S., Skipp, A., & Haywood, S. (2015). Durham Shared Maths Project. In *Education Endowment Foundation.* https://educationendowmentfoundation.org.uk/uploads/pdf/Shared_Maths_1.pdf

Lord, P., Rabiasz, A., Styles, B., Andrade, J. (2017). *"Literacy Octopus" Dissemination Trial*.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned? *Educational Researcher*, *48*(3), 158–166.

https://doi.org/10.3102/0013189X19832850

Lysenko, L. V, Abrami, P. C., Bernard, R. M., Dagenais, C., & Janosz, M. (2014). Educational research in educational practice: Predictors of use. *Canadian Journal of Education*, *37*(2), 1–26.

Mahon, W.A., Daniel, E. E. (1964). A method for the assessment of reports of drug trials. *Can Med Assoc J*, *90*, 565–569.

Makel, M.C., Plucker, J. A. (2014). Creativity is more than novelty: Reconsidering replication as a creativity act. *Psychology of Aesthetics, Creativity, and the Arts.*, *8*(1), 27–29.

Malički, M., & Marušić, A. (2014). Is there a solution to publication bias? Researchers call for changes in dissemination of clinical research results. *Journal of Clinical Epidemiology*, *67*(10), 1103–1110. https://doi.org/10.1016/j.jclinepi.2014.06.002

Mason, J. (1994). Linking qualitative and quantitative data analysis. In *Analysing Qualitative Data* (pp. 89–110).

Mason, J. (1996). *Qualititative Researching*. Sage Publications.

Mathes, K.L., & Fuchs, L. S. (1994). The Efficacy of peer tutoring in reading for students with mild disabilities: A best-evidence synthesis. *School Psychology Review*, *23*, 59–80.

Mayou, R. . (2000). Psychological debriefing for road traffic accident victims: Three-year follow-up of a randomised controlled trial. *Journal The British Journal of Psychiatry*, *176*(6), 578.

McCord, J. (1978). A thirty-year follow-up of treatment effects. *American Psychologist*, *33*, 284–289.

McCord, J. (2001). Crime prevention: a cautionary tale. *3rd Evidence-Based Policies and Indicator Systems Conference.*

McMillan, J.H., Wergin, J. F. (2006). *Understanding and Evaluating Educational Research*. Pearson.

Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). Evaluation of Evidence-Based Practices in Online Learning. In *U.S. Department of Education*. www.ed.gov/about/offices/list/opepd/ppss/reports.html

Measor, L., & Woods, P. (1984). *Changing schools: Pupil perspectives on transfer to a comprehensive.* Milton Keynes: Open University Press.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology*, *49*(2), 270–291. https://doi.org/10.1037/a0028228

Mhaskar, R., Djulbegovic, B., Magazin, A., Soares, H. P., & Kumar, A. (2012). Published methodological quality of randomized controlled trials does not reflect the actual quality assessed in protocols. *Journal of Clinical Epidemiology*, *65*(6), 602–609. https://doi.org/10.1016/j.jclinepi.2011.10.016

Mill, J. (1882). *A System of Logic, Ratiocinative and Inductive, Being a Connected View of Principles of Evidence, and the Methods of Scientific Investigation (8th Edition)*. Harper & Brothers.

Mitchell, N., Ainsworth, H., Buckley, H., Hewitt, C., Torgerson, D., & Torgerson, C. (2016). Pragmatic cluster randomised controlled trial of contextualised grammar teaching and small group teaching to improve the writing skills of 11 year old children. *OnlineEduactional Research Journal*.

Mitton, C., Adair, C. E., McKenzie, E., Patten, S. B., & Perry, B. W. (2007). Knowledge Transfer and Exchange: Review and Synthesis of the Literature. *Milbank Quarterly*, *85*(4), 729–768. https://doi.org/10.1111/j.1468-0009.2007.00506.x

Moher, D., Dulberg C.S., W. G. A. (1994). Statistical power, sample size and their reporting in

randomized controlled trials. *Journal of the American Medical Association*, *272*(2), 122–124.

Moher, D., Schuluz, K.F., Altman, D. (2001). The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomised Trials. *JAMA*, *285*(15), 1987–1991. http://www.consort-statement.org/consort-statement/overview0/

Montgomery, P., Grant, S., Mayo-Wilson, E., Macdonald, G., Michie, S., Hopewell, S., Moher, D., Lawrence Aber, J., Altman, D., Bhui, K., Booth, A., Clark, D., Craig, P., Eisner, M., Sherman, L., Fraser, M. W., Gardner, F., Hedges, L., Hollon, S., … Yaffe, J. (2018). Reporting randomised trials of social and psychological interventions: The CONSORT-SPI 2018 Extension. *Trials*, *19*(1), 1–14. https://doi.org/10.1186/s13063-018-2733-1

Moran, J., Ferdig, R., Pearson, P. D., Wardrop, J., & Blomeyer, R. (2008). Technology and Reading Performance in the Middle-School Grades: A Meta-Analysis with Recommendations for Policy and Practice. *Journal of Literacy Research*, *40*(1), 6–58. https://doi.org/10.1080/10862960802070483

Morgan, D. L. (1996). Focus Groups. *Annual Review of Sociology*, *22*, 127–152.

Morrison, K. (2001). Randomised Controlled Trials for Evidence-based Education: Some Problems in Judging "What Works." *Evaluation & Research in Education*, *15*(2), 69–83. https://doi.org/10.1080/09500790108666984

Mowat, J. G. (2019). Supporting the transition from Primary to Secondary school for pupils with social, emotional and behavioural needs: a focus on the socio-emotional aspects of a transfer for an adolescent boy. *Emotional and Behavioural Difficulties*, *24*(1), 50–69.

MRC. (2000). *Developing and evaluating complex interventions*.

National Audit Office. (2015). *Funding for disadvantaged pupils summary* (Issue June). https://doi.org/DP Ref: 10761-001

Nelson, J. & Campbell, C. (2017) Evidence-informed practice in education: meanings and applications, *Educational Research*, 59 (2), 127-135.

Nelson, J., & O'Beime, C. (2014). *Using Evidence in the classroom: What Works and Why? Research Summary*.

Nelson, Julie, & O'Beirne, C. (2014). *Using evidence in the classroom: What works and why? Report*. https://www.nfer.ac.uk/publications/IMPA01/IMPA01.pdf

Noffke, S. E. (1992). The work and workplace of teachers in action research. *Teaching & Teacher Education*, *8*(1), 15–29.

Norris, N. (1996). Professor Hargreaves, the TTA and evidence-based practice. *Research Intelligence*, Number 57, 2-4.

Nutley, S.M., Percy-Smith, J., Solesbury, W. (2003). *Models of research impact: A cross-sector review of literature and practice: Building effective research.*

Nutley, S., Walter, I., & Davies, H. T. O. (2009). Promoting evidence-based practice: Models and mechanisms from cross-sector review. *Research on Social Work Practice*, *19*(5), 552–559. https://doi.org/10.1177/1049731509335496

Ofsted. (2015). *Key Stage 3: The Wasted Years*.

Oztok, M., Zingaro, D., Brett, C., & Hewitt, J. (2013). Exploring asynchronous and synchronous tool use in online courses. *Computers and Education*, *60*(1), 87–94. https://doi.org/10.1016/j.compedu.2012.08.007

Peña, J., & Hancock, J. T. (2006). An Analysis of Socioemotional. *Communication Research*, 92–109. https://doi.org/10.1177/0093650205283103

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-Reporting Bias in Education Research. *Educational Researcher*, *42*(October), 424–432. https://doi.org/10.3102/0013189X13507104

Pildal, J., Chan, A.-W., Hróbjartsson, A., Forfang, E., Altman, D. G., & Gøtzsche, P. C. (2005). Comparison of descriptions of allocation concealment in trial protocols and the published reports: cohort study. *BMJ (Clinical Research Ed.)*, *330*(April), 1049. https://doi.org/10.1136/bmj.38414.422650.8F

Pocock, S.J., Hughes, M.D., Lee, R. J. (1987). Statistical problems in reporting clinical trials. *N Engl J Med*, *317*, 426–432.

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. a. (2015). Estimating the Difference Between Published and Unpublished Effect Sizes: A Meta-Review. *Review of Educational Research*, *1975*, 1–30. https://doi.org/10.3102/0034654315582067

Proctor, R. (2015). Teachers and school research practices: the gaps between the values and practices of teachers. *Journal of Education for Teaching*, *41*(5), 464–477.

Ritter, G. W., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The Effectiveness of Volunteer Tutoring Programs for Elementary and Middle School Students: A Meta-Analysis. *Review of Educational Research*, *79*(1), 3–38. https://doi.org/10.3102/0034654308325690

Rockinson-Szapkiw, A. (2009). The Impact Of Asynchronous And Synchronous Instruction And Discussion On Cognitive Presence, Social Presence, Teaching Presence, And Learning. *Ph.D. Thesis Regent University*.

Rogers, E. (1983). *Diffusion of Innovation* (3rd Editio). Free Press.

Rogers, E. (2003). *Diffusion of innovations* (5th Editio). Simon & Schuster International.

Rohrbeck, C. a., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology*, *95*(2), 240–257. https://doi.org/10.1037/0022-0663.95.2.240

Rose, J., Thomas, S., Zhang, L., Edwards, A., Augero, A., Roney, P. (2017). *Research Learning Communities*.

Rosenthal, R. (1980). Combining probabilities and the file drawer problem. *Evaluation in Education*, *4*, 18–21. https://doi.org/10.1016/0191-765X(80)90003-8

Rothstein, H.R., Sutton, A.J., & Borenstein, M. (eds). (2005). *Publication bias in Meta-Analysis: prevention, assessment and adjustments*. John Wiley & Sons LTD.

Rourke, L., & Kanuka, H. (2009). Learning in communities of inquiry: a review of the literature. *Journal of Distance Education*, *23*, 19–48.

Ruane, R. (2012). *A Study of Student Interaction in an Online Learning Environment Specially Crafted for Cross-Level Peer Mentoring A Thesis Submitted to the Faculty of Drexel University by Regina Ruane in partial fulfillment of the requirements for the degree of Doctor of* (Issue August) [Drexel University]. https://doi.org/ProQuest Disssertations: 3533696

Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C.K., Torgesen, J. K. (2007). *Reading interventions for adolescent struggling readers: A meta-analysis with implications for practice*.

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2013). A Meta-Analysis of Interventions

for Struggling Readers in Grades 4-12: 1980-2011. *Journal of Learning Disabilities*, 0022219413504995. https://doi.org/10.1177/0022219413504995

Schulz, K. F, Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Bmj*, *340*(mar23 1), c332–c332. https://doi.org/10.1136/bmj.c332

Schulz, Kenneth F, Altman, D. G., Moher, D., & Group, C. (2010). Academia and Clinic Annals of Internal Medicine CONSORT 2010 Statement : Updated Guidelines for Reporting Parallel Group Randomized Trials OF TO. *Annals of Internal Medicine*, *1996*(14), 727–732. https://doi.org/10.7326/0003-4819-152-11-201006010-00232

Schulz, Kenneth F, & Grimes, D. A. (2013). Get in the spirit with SPIRIT 2013: protocol content guideline for clinical trials. *Contraception*, *88*(6), 676–677. https://doi.org/10.1016/j.contraception.2013.10.002

Schwier, R. A., & Balbar, S. (2002). The interplay of content and community in synchronous and asynchronous communication: Virtual communication in a graduate seminar. *Canadian Journal of Learning and Technology*, *28*.

See, B. H. (2018). Evaluating the evidence in evidence-based policy and practice: Examples from systematic reviews of literature. *Research in Education*, *102*(1), 37–61.

Shadish, W.R., Cook, T.D., Campbell, D. T. (2002). *Experimental and Designs for Generalized Causal Inference*.

Sharples, J. (2013). *Evidence for the frontline: A report for the alliance for useful evidence* (Issue June). http://www.alliance4usefulevidence.org/publication/evidence-for-the-frontline/

Shepherd, J., & Roker, D. (2005). *An evaluation of a 'transition to secondary school' project run by the National Pyramid Trust.* Brighton: Trust for the Study of Adolescence.

Shepherd, J. (2007). The production and management of evidence for public service reform. *Evidence & Policy: A Journal of Research, Debate & Practice*, *3*(2), 231–251. http://0-search.ebscohost.com.brum.beds.ac.uk/login.aspx?direct=true&db=sih&AN=26316574&site=ehost-live&scope=site

Sibieta, L. (2018). *School funding per pupil falls faster in England than in Wales*. Institue Fiscal Studies. https://www.ifs.org.uk/publications/13143

Siddiqui, N., Gorard, S., & See, B. H. (2015). Accelerated Reader as a literacy catch-up intervention during primary to secondary school transition phase. *Educational Review*, *1911*(November), 1–16. https://doi.org/10.1080/00131911.2015.1067883

Simes, R. J. (1986). Publication bias: The case for an international registry of clinical trials. *Jouranl of Clinical Oncology*, *4*, 1529–1541.

Slavin, R.E., Cheung, A., Groff, C., Lake, C. (2008). Effective Reading Programs for Middle and High Schools : A Best Evidence Synthesis. *Reading Research Quarterly*, *43*(3), 290–322.

Slavin, R. . (2002). Evidence-Based Education Policies : Transforming Educational Practice and Research. *Educational Researcher*, *31*(7), 15–21.

Slavin, R. . (2004). Educational Research Can and Must Adress "What Works " Questions. *Educational Researcher*, *33*(February 2004), 27–28.

Slavin, R. E., Hanley, P., Lake, C., & Hanley, P. (2012). The Effective Programs for Elementary Science : A Best-Evidence Synthesis Effective Programs for Elementary Science : A Best-Evidence Synthesis University of York. *Best Evidence Encyclopaedia*, *April 2016*.

Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*(1), 1–26. https://doi.org/10.1016/j.edurev.2010.07.002

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective Programs in Middle and High School Mathematics: A Best-Evidence Synthesis. *Review of Educational Research*, *79*(2), 839–911. https://doi.org/10.3102/0034654308330968

Slavin, R., & Smith, D. (2009). The Relationship between Sample Sizes and Effect Sizes in Systematic Reviews in Education. *Educational Evaluation and Policy Analysis*, *31*(4), 500–506. https://doi.org/10.3102/0162373709352369

Smith, M. (1980). Publication bias and Meta-analysis. *Evaluation in Education*, *4*, 22–24.

Smyth, R. M. D., Kirkham, J. J., Jacoby, A., Altman, D. G., Gamble, C., & Williamson, P. R. (2011). Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists. *BMJ (Clinical Research Ed.)*, *342*, c7153. https://doi.org/10.1136/bmj.c7153

Speight, S., Callanan, M., Griggs, J., & Farias, J. C. (2016). *Rochdale Research into Practice*.

Stampfer, M.J., Goldhaber, S.V., Usuf, S., Peto, R., & Henekens, C. H. (1982). Effect of intravenous streptokinase on acute myocardial infarction: pooled results from randomised trials. *New England Journal Medicine*, *307*, 1180–1182.

Sterling, T. D. (1959). Publication Descisions and Their Possible Effects on Inferences Drawn from Tests of Significance -- Or Visa Versa. *Journal of the American Statistical Association*, *54*(285), 30–34.

Styles, B., Clarkson, R., Fowler, K. (2014). *Chatterbooks: Evaluation Report and Executive Summary*.

Styles, B., Galvis, M., Lord, P. Roy, P. (2019). *Teacher Choices Trial: A winning start.*

Tetzlaff, J. M., Chan, A.-W., Kitchen, J., Sampson, M., Tricco, A. C., & Moher, D. (2012a). Guidelines for randomized clinical trial protocol content: a systematic review. *Systematic Reviews*, *1*(1), 43. https://doi.org/10.1186/2046-4053-1-43

Tetzlaff, J. M., Chan, A.-W., Kitchen, J., Sampson, M., Tricco, A. C., & Moher, D. (2012b). Guidelines for randomized clinical trial protocol content: a systematic review. *Systematic Reviews*, *1*, 43. https://doi.org/10.1186/2046-4053-1-43

The Lancet. (2010). Strengthening the credibility of clinical research. *Lancet*, *375*(9722), 1225. https://doi.org/10.1016/S0140-6736(10)60523-5

Thurston, A., & Topping, K. J. (2008). A randomised trial of paired reading in elementary schools. *American Educational Research Association*.

Thurston, Allen, Duran, D., Cunningham, E., Blanch, S., & Topping, K. (2009). International on-line reciprocal peer tutoring to promote modern language development in primary schools. *Computers and Education*, *53*(2), 462–472. https://doi.org/10.1016/j.compedu.2009.03.005

Thurston, Allen, Miller, S., Dunne, L., Lazenbatt, A., Gildea, A., Stepien, D., & Tapsell, D. (2015). Research protocol: A randomized controlled trial of functional family therapy: An Early Intervention Foundation (EIF) partnership between Croydon Council and Queen's University Belfast. *International Journal of Educational Research*, *70*, 47–56. https://doi.org/10.1016/j.ijer.2015.01.001

Tonks, A. (2002). A clinical trials register for Europe. *Bmj*, *325*(December), 1314–1315. https://doi.org/10.1136/bmj.325.7376.1314

Tooley, J., Darby, D. (1998). *Educational Research: A critique*.

Topping, K., Campbell, J. (2003). Cross-age peer tutoring in mathematics with seven and eleven year olds: Influence on mathematical vocabulary, dialogue and self-concept. *Educational Research*, *45*, 287–308.

Topping, K. J., Dehkinet, R., Blanch, S., Corcelles, M., & Duran, D. (2013). Paradoxical effects of feedback in international online reciprocal peer tutoring. *Computers and Education*, *61*(1), 225–231. https://doi.org/10.1016/j.compedu.2012.10.002

Topping, K., Miller, D., Murray, P., & Conlin, N. (2011). Implementation integrity in peer tutoring of mathematics. *Educational Psychology*, *31*(5), 575–593. https://doi.org/10.1080/01443410.2011.585949

Torgerson, C., Wiggins, A., Torgerson, D., Ainsworth, H., Hewitt, C. (2013). Every Child Counts: testing policy effectiveness using a randomised controlled trial, designed, conducted and reported to CONSORT standards. *Research in Mathematics Education*, *15*(2), 141–153.

Torgerson, C.J., Wiggins, A., Torgerson, D.J., Ainsworth, H., Barmby, P., Hewitt, C., Jones, K., Hendry, V., Askew, M., Bland, M., Coe, R., Higgins, S., Hodgen, J., Hulme, C. & Tymms, P. (2011). *Every Child Counts: the independent evaluation technical report*.

Torgerson, D., Torgerson, C. (2014). Grammar for Writing. *Education Endowment Foundation.* https://educationendowmentfoundation.org.uk/uploads/pdf/FINAL_EEF_Evaluation_Report_-_Grammar_for_Writing_-_February_2014.pdf

Torgerson, C., Ainsworth, H., Buckley, H., Hampden-Thompson, G., Hewitt, C., Humphry, D., Jefferson, L., Mitchell, N., & Torgerson, D. (2016). *Affordable Online Maths Tuition* (Issue July).

Torgerson, C.J. (2009). Randomised controlled trials in education research: a case study of an individually randomised pragmatic trial. *Education 3-13*, *37*(4), 313–321. https://doi.org/10.1080/03004270903099918

Torgerson, C J, & Torgerson, D. J. (2001). *The need for randomised controlled trials in educational research*. *49*(3), 316–328. https://doi.org/10.1111/1467-8527.t01-1-00178

Torgerson, C J, Torgerson, D. J., Birks, Y. F., & Porthouse, J. (2005). A comparison of randomised controlled trials in health and education. *British Educational Research Journal*, *31*(6), 761–785. https://doi.org/10.1080/01411920500314919

Torgerson, Carole J. (2006). Publication Bias: the Achilles' Heel of Systematic Reviews? *British Journal of Educational Studies*, *54*(1), 89–102. https://doi.org/10.1111/j.1467-8527.2006.00332.x

Trochim, W. (2006). *The Research Methods Knowledge Base.* http://www.socialresearchmethods.net/kb/citing.php

Trust, S. (2017). *SMALL BUT INCREASING NUMBER OF SCHOOLS ARE USING THEIR PUPIL PREMIUM FUNDING TO OFFSET BUDGET CUTS NEW SUTTON TRUST AND EDUCATION ENDOWMENT FOUNDATION POLLING*.

Tsuei, M. (2011). Development of a peer-assisted learning strategy in computer-supported collaborative learning environments for elementary school students. *British Journal of Educational Technology*, *42*(2), 214–232. https://doi.org/10.1111/j.1467-8535.2009.01006.x

Tsuei, M. (2012). Using synchronous peer tutoring system to promote elementary students' learning in mathematics. *Computers and Education*, *58*(4), 1171–1182. https://doi.org/10.1016/j.compedu.2011.11.025

Tu, C. H., & Corry, M. (2003). Designs, management tactics, and strategies in asynchronous learning discussions. *The Quarterly Review of Distance Education*, *4*, 303–315.

Tymms, P., Merrell, C., Thurston, a., Andor, J., Topping, K., & Miller, D. (2011). *Improving attainment across a whole district : school reform through peer tutoring in a randomized controlled trial. October*, 37–41. https://doi.org/10.1080/09243453.2011.589859

University Press, O. (2016). *Oxford Online Dictionary*. http://www.oxforddictionaries.com/definition/english/protocol

Vulliamy, G. (1998). A complete misubderstanding of our position... trivalises our arguments. *Research Intelligence*, 17–18.

Walter, I., Nutley, S., Percy-Smith, J., McNeish, D., & Frost, S. (2004). *Knowledge Review 7: Improving the use of research in social care practice*.

Wessely, S., Rose, S., Bisson, J. (1998). A systematic review of brief psychological interventions ("debriefing") for the treatment of immediate trauma related symptoms and the prevention of post traumatic stress disorder. In *Cochrane Libary* (Vol. 4).

West, P., Sweeting, H., & Young, R. (2010). Transition matters: pupils' experiences of the primary-secondary school transition in the West of Scotland and consequences for well-being and attainment. In *Research Papers in Education* (Vol. 25, Issue 1). https://doi.org/10.1080/02671520802308677

Wiggins, M., Jerrim, J., Tripney, J., Khatwa, M., Gough, D. (2019). *The RISE Project: Evidence-informed school improvement*.

Wiglesworth, M., Lendrum, A., Oldfield, J., Scott, A., ten Bokkel, I., Tate, K., & Emery, C. (2016). The impact of trial stage, developer involvement and international transferability on universal social and emotional learning progrmme outcomes: a meta-analysis. *Cambridge Journal of Education*, *46*(3), 347–376.

Williams, D., & Coles, L. (2007). Teachers' approaches to finding and using research evidence: an information literacy perspective. *Educational Research*, *49*(2), 185–206. https://doi.org/10.1080/00131880701369719

Wilson, T., & Walsh, C. (1996). *Information behaviour: an interdisciplinary perspective*. British Library Reserach and Innovative Report (10). http://www.informationr.net/tdw/publ/infbehav/

Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. (1994). Call for comments on a proposal to improve reporting of clinical trials in the biomedical literature. Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. *Ann Intern Med.*, *121*(11), 894–895.

Yan, B., Peng, L., Zhao, X., Chung, H., Li, L., Zeng, L., Ong, H., & Wang, G. (2014). Nesiritide fails to reduce the mortality of patients with acute decompensated heart failure : An updated systematic review and cumulative meta-analysis. *International Journal of Cardiology*, *177*(2), 505–509. https://doi.org/10.1016/j.ijcard.2014.08.078

Zeedyk, M. S., Gallacher, J., Henderson, M., Hope, G., Husband, B., & Lindsay, K. (2003). Negotiating the Transition from Primary to Secondary School: Perceptions of Pupils, Parents and Teachers. *School Psychology International*, *24*(1), 67–79. https://doi.org/10.1177/0143034303024001010

Zeneli, M., Thurston, A., & Roseth, C. (2016). The influence of experimental design on the magnitude of the effect size -peer tutoring for elementary, middle and high school settings: A meta-analysis. *International Journal of Educational Research*, *March*. https://doi.org/10.1016/j.ijer.2015.11.010

Zeuli, J. S. (1994). How do teachers understand research? *Teaching & Teacher Education*, *10*(1), 39–55. https://doi.org/10.1016/0742-051X(94)90039-6

Zhang, C., Wang, Z.-Y., Qin, Y.-Y., Yu, F.-F., & Zhou, Y.-H. (2014). Association between B vitamins supplementation and risk of cardiovascular outcomes: a cumulative meta-analysis of randomized controlled trials. *PloS One*, *9*(9), e107060. https://doi.org/10.1371/journal.pone.0107060

Appendices

Appendix 1 – Early Stage Protocol Study One

Durham University

School of Education

Research Ethics and Data Protection Monitoring Form

Research involving humans by all academic and related Staff and Students in the Department is subject to the standards set out in the Department Code of Practice on Research Ethics. The Sub-Committee will assess the research against the British Educational Research Association's *Revised Ethical Guidelines for Educational Research* (2011).

It is a requirement that prior to the commencement of all research this form be completed and submitted to the Department's Research Ethics and Data Protection Sub-Committee. The Committee will be responsible for issuing certification that the research meets ethical standards and will, if necessary, require changes to the research methodology or reporting strategy.

The application should contain:

1) this completed (and signed) application form
2) a copy of the research proposal which should be no longer than one A4 page that details: (a) objectives of the study, (b) description of the target cohort / sample (c) methods and procedure of data collection, (d) data management and (f) reporting strategies.
3) depending on the methods you plan to employ, outline of the interview schedule / survey / questionnaire / or other assessment methods
4) the participant information sheet, and
5) the consent form

Templates for the participant information sheet and the consent form are provided below.

Incomplete applications will be returned without consideration.

Please send all documents to the Research Office in the School of Education (Sheena Smith, School of Education, tel. (0191) 334 8403, e-mail: Sheena.Smith@Durham.ac.uk).

Application for Ethics Approval

Name	Wayne Harrison
Email address	w.d.harrison@durham.ac.uk
Title of research project	Online Maths Peer Tuition – Pilot study
Date of start of research project	November 2015

	Please tick one		
PGR Student	X	*For PGR, PGT and UG students*	
PGT Student		Programme	PhD Education
UG Student		Supervisor	Prof Steve Higgins
		For staff	
Staff		Is the research funded	N
		Funder	
		List any Co-Is in the research	
		Other	
Other		Please give further details	

(1) Does the proposed research project involve data from human subjects? This includes secondary data. (1) Y

If the research project is concerned with the analyses of secondary data (e.g. pre-existing data or information records) please continue with Q6-9

(2) Will you provide your informants – prior to their participation – with a participant information sheet containing information about

(2a) the purpose of your research (2a) Y

(2b) the voluntary nature of their participation (2b) Y

(2c) their right to withdraw from the study at any time (2c) Y

(2d) what their participation entails (2d) Y

(2e) how anonymity is achieved (2e) Y

(2f) how confidentiality is secured	(2f)	Y
(2g) whom to contact in case of questions or concerns	(2g)	Y
Please attach a copy of the information sheet or provide details of alternative approach.		
(3) Will you ask your informants to sign an informed consent form?	(3)	N
(please attach a copy of the consent form or provide details of alternative approach)	Alternative added	
(4) Does your research involve covert surveillance?	(4)	N
(4a) If yes, will you seek signed consent post hoc?	(4a)	
(5) Will your data collection involve the use of recording devices?	(5)	N
(5a) If yes, will you seek signed consent?	(5a)	
(6) Will your research report be available to informants and the general public without restrictions placed by sponsoring authorities?	(6)	Y

(7) How will you guarantee confidentiality and anonymity?

The analysis will be completed on an anonymous data set. The data will be password protected and not shared with any other people.

(8) What are the implications of your research for your informants?

The study is designed to help support students identified as more able and talented in Mathematics and the study will hopefully have positive implications in terms of extending knowledge and attainment.

(9) Are there any other ethical issues arising from your research?

The risks of harm are incredibly small, as the students will be completing normal educational activities. I am not using a control group so all students are receiving the same intervention, with only the group size differing. The use of an active control ensures all pupils receive the intervention.

Further details

I have decided to use an opt out consent form as a number of high profile trials running similar projects for the EFF use this format. Please see the York EEF evaluation of the Nesta / Third space evaluation protocol design by Prof David and Carole Torgerson.

http://educationendowmentfoundation.org.uk/uploads/pdf/Online_Maths_Tutoring.pdf

As my project is involving 5 schools I feel that the opt out consent form with a detailed parental / student letter is sufficient for the pilot study. I have included these with my ethics proposal.

Declaration

I have read the Department's Code of Practice on Research Ethics and believe that my research complies fully with its precepts.

I will not deviate from the methodology or reporting strategy without further permission from the Department's Research Ethics Committee.

I am aware that it is my responsibility to seek and gain ethics approval from the organisation in which data collection takes place (e.g., school) prior to commencing data collection.

Signed ……W.Harrison…….. Date 14.9.15…

Proposal discussed and agreed by supervisor (for students) or colleague (for staff):

Name Prof Steve Higgins……. On …(Date)

[DATE]

Participant Information Sheet

Title:

You are invited to take part in a research study of online peer to peer tuition in Mathematics. Please read this form carefully and ask any questions you may have before agreeing to be in the study.

The study is conducted by Wayne Harrison as part of his MA studies at Durham University. This research project is supervised by Prof Steve Higgins (s.e.higgins@durham.ac.uk) from the School of Education at Durham University.

The purpose of this study is to investigate if online one to many and one to two peer tutoring can be as effective as one to one online peer tuition in Mathematics.

If you agree to be in this study, you will be asked to complete a short pre and post-test and complete 10 online tutorials.

Your participation in this study will take approximately 300 minutes.

You are free to decide whether or not to participate. If you decide to participate, you are free to withdraw at any time without any negative consequences for you.

All responses you give or other data collected will be kept confidential. The records of this study will be kept secure and private. All files containing any information you give are password protected. In any research report that may be published, no information will be included that will make it possible to identify you individually. There will be no way to connect your name to your responses at any time during or after the study.

* FUNDING (explain where FUNDING for this project comes from)

If you have any questions, requests or concerns regarding this research, please contact me via email at w.d.harrison@durham.ac.uk or by telephone at [TELEPHONE NUMBER].

This study has been reviewed and approved by the School of Education Ethics Sub-Committee at Durham University (date of approval: DD/MM/YY)

Declaration of Informed Consent

- I agree to participate in this study, the purpose of which is to [GENERIC DESCRIPTION OF THE PURPOSE OF THE STUDY].
- I have read the participant information sheet and understand the information provided.
- I have been informed that I may decline to answer any questions or withdraw from the study without penalty of any kind.
- I have been informed that all of my responses will be kept confidential and secure, and that I will not be identified in any report or other publication resulting from this research.
- I have been informed that the investigator will answer any questions regarding the study and its procedures. [NAME OF RESEARCHER], School of Education, Durham University can be contacted via email: [DURHAM EMAIL ADDRESS OF RESEARCHER] or telephone: [TELEPHONE NUMBER].
- I will be provided with a copy of this form for my records.

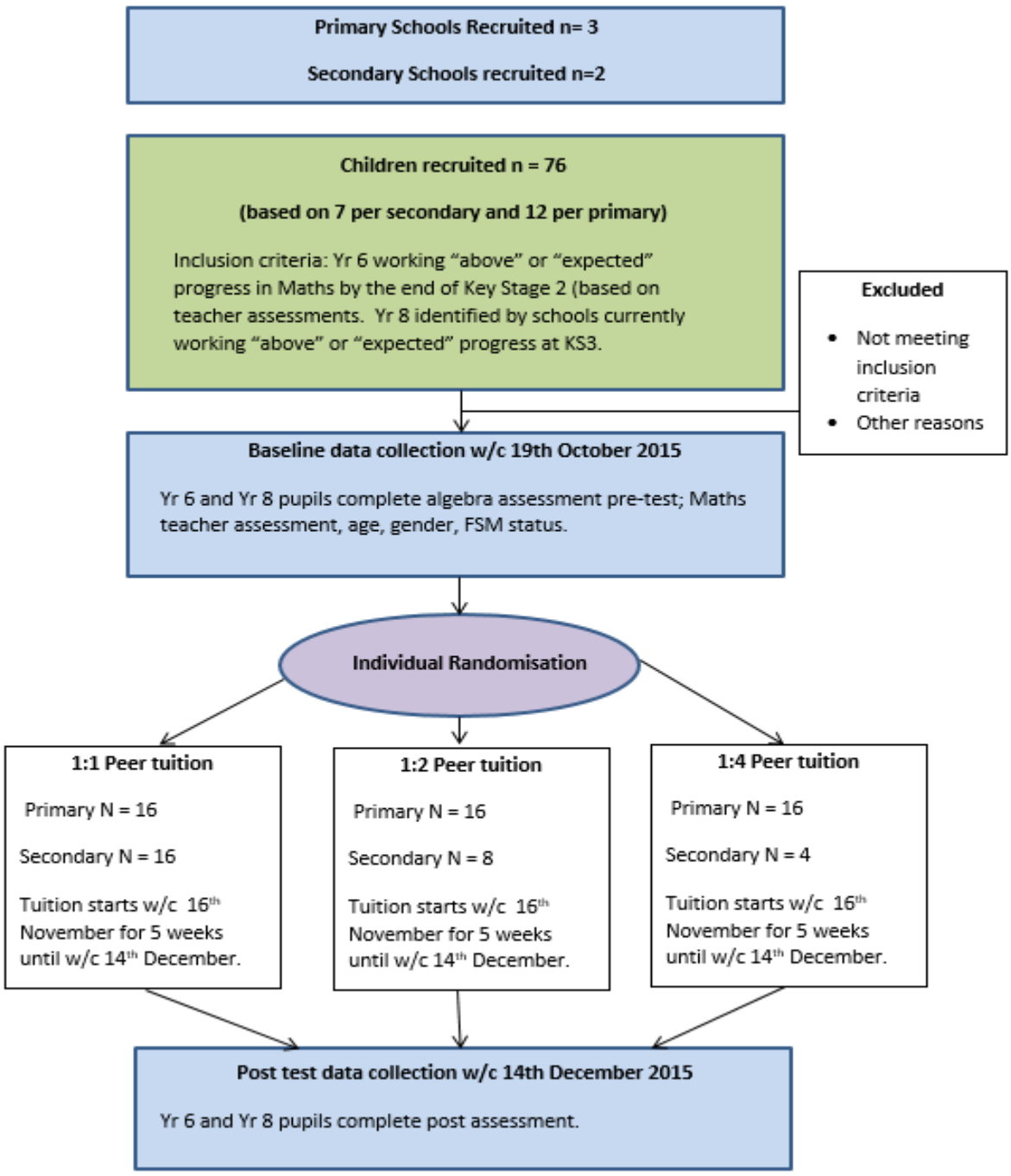
Any concerns about this study should be addressed to the Ethics Sub-Committee of the School of Education, Durham University via email (Sheena Smith, School of Education, tel. (0191) 334 8403, e-mail: Sheena.Smith@Durham.ac.uk).

Date Participant Name (please print) Participant Signature

I certify that I have presented the above information to the participant and secured his or her consent.

Date Signature of Investigator

Trial diagram



Protocol

Pilot efficacy study for small scale aggregated trials investigating the effectiveness of online cross-age peer tuition group sizes across the transition boundary between primary and secondary schools

Version: 2.0

Date: December 2015

Study lead: Wayne Harrison (Durham University)

Contact: w.d.harrison@durham.ac.uk

Background and significance

The study is funded as part of an ESRC funded PhD at Durham University to investigate the feasibility of online cross-age peer tuition between primary and secondary schools. The study is developing a new methodology for aggregating scale small trials into a prospective cumulative meta-analysis.

The intervention will provide primary pupils in Year 6 online peer tuition in the subject of mathematics, using Year 8 secondary pupils to peer teach in group sizes of 1:1, 1:2 and 1:4. The online peer tuition lessons are provided in advance to eh Year 8 pupils and they deliver these in weekly 30 minute lessons via an online platform.

The intervention aims to help improve attainment in mathematics for the topic of data handling while they are in their final year of their studies.

Research question

The trial aims to determine the feasibility of implementing a more able and talented cross-age peer tuition intervention across the transition boundary between primary and secondary schools.

Primary Objective

The primary research questions is:

1) Is the implementation of an online cross-age peer tuition intervention feasible between primary and secondary schools as an aggregated trial?
    a) Is online cross-age peer tuition a feasible intervention strategy between schools as part of an aggregated trial?
    b) Are the instruments used appropriate for the ability of students involved in the study?
    c) Can the intervention maintain fidelity if delivered to schools in other locations in the UK?

Secondary Objectives

Secondary objectives are:

1) Does one-to-two online cross-age peer-tutoring produce equivalent effects in mathematics attainment[24] compared with online one-to-one tuition?
2) Does one-to-many online cross-age peer-tutoring produce equivalent effects in mathematics attainment[2] as compared with one-to-one tuition?

Design

The design for the study will be a pragmatic randomised controlled trial using an aggregation of three smaller trials over a sequential timescale of 9 months. The individual trials will be conducted as an efficacy study as the intervention will be required to be conducted under highly controlled and optimal conditions in order to maximise outcomes, and ensure the technology worked in schools.

Recruitment

Recruitment will target secondary schools with feeder primary schools in the North East of England, recruiting 7 Year 8 peer tutors (12 – 13 years) to tutor 12 Year 6 pupils (10-11 year) per school partnership. The research lead will provide an information document for parents and pupils and it is anticipated that the intervention will be delivered in Autumn term 2, Spring term and Summer terms for the three cohorts of schools.

Inclusion criteria

Each participating secondary school will identify 7 pupils and each primary school identified 12 pupils who are eligible to participate in the trial. Since the removal of the national framework of 'levels' to report pupils attainment and progress it was not possible to specify specific inclusion criteria. A relative criterion, depending on each individual school assessment – "above expected", "expected" and "below expected" progress will be used. Consequently, consistency between individual schools is a potential limitation.

The inclusion criteria that will be used in the study are:

[24] Mathematics attainment is defined as mastery level concepts in KS2 curriculum for Algebra.

School inclusion criteria: Secondary and primary schools are eligible to take part in the trial if they agree to all trial procedures including provision of pupil data, informing parents, randomisation and implementation of the intervention.

Primary pupil inclusion criteria: Pupils eligible for the intervention will be either at "expected" or "above expected" for progress in Mathematics by the end of Year 6, based on teacher assessments.

Secondary pupil inclusion criteria: Pupils eligible for the intervention will be either at "expected" or "above expected" for progress in Mathematics by the end of Year 8, based on teacher assessments.

Each school will provide parents with an information letter and an optional opt out form if they do not wish their child to participate in the research. Please see the appendix for examples of these documents.

School participation

Schools are required to sign a memorandum of understanding which outlines the following:

- I confirm I have read and understood the information sheet for the above project and have had the opportunity to ask questions;
- I understand that all children's results will be kept confidential and protected using encryption software and that no material which could identify individual children or the school will be used in any reports of this project;
- I agree for the pupils to undertake an assessment pre and post completion of the project;
- I agree for the pupils to complete assessments under exam conditions at agreed dates / time;
- I agree to providing the information letter to all parents of children involved in the project;
- I consent to Wayne Harrison from Durham University having access to the pupil data to be used solely to evaluate the impact of the project;

Intervention

The intervention is an online mathematics programme designed for weekly 30-minute online peer tutoring for the data handling topic.

Randomisation

The Year 8 pupils will be randomised using unequal randomisation, controlling for any unknown hidden biases when conducting the final analyses. The Year 6 pupils will be block randomised using pre-test data ensuring the active control and two comparator groups are balanced for the mathematics ability as defined by the data handling assessment. The design is ethical as all pupils are participating in the intervention, with only the group sizes changing for the delivery of the peer tuition in group sizes 1:1, 1:2 and 1:4.

Sample size calculation

The power calculations are based upon the following. As the intervention is comparing the group sizes 1:2 and 1:4 with the active control 1:1, it is reasonable to conclude that the differences are likely to be smaller than the difference between a standard business as a usual control[25].

It is important to note that the pilot study will not be sufficiently powered from the three cohorts of schools in this study (Kraemer & Blasey, 2016). The following power calculation provides a theoretical sample size for the aggregated trials to reach a sufficient sample size to allow generalisations to be inferred (and to balance the risks of Type 1 and Type 2 errors).

The power calculations are based on the following assumptions:

5) A school-level intra-cluster correlation (ICC) of 0.15
6) Equal cluster sizes of 4 per school (3 groups of 4 per school)
7) 30 Schools (clusters) participating in the aggregated trial
8) 80% power for 95% confidence interval

For the purposes of calculating the sample size it is assumed that 30 schools (online tuition groups) would be recruited over the duration of the aggregated trial with 12 pupils per primary school; this would result in a total sample size of 360 pupils. Assuming 4 pupils per school were allocated to the 1:1 group and an intra-cluster correlation coefficient of 0.15, based upon the assumptions, the trial would be able to detect an effect of 0.19 standard deviations with 80% certainty (Kraemer & Blasey, 2016).

[25] Group A (1:1) is the active control and the comparisons are B (1:2) vs A and C (1:4) vs A. Therefore it is not required to take into account any increased type 1 error.

Outcome measures

The outcome measures will use the scores from the mathematics assessment used as he pre- and post-assessment for the research.

Analysis

Impacts will be estimated on the basis of intention to treat, whereby all pupils who are initially involved in the pre – and post-testing will be analysed according to the group they were initially assigned, regardless of whether they went on to participate in the intervention.

Effect sizes will be calculated and presented alongside their 95% confidence intervals. The effect size reported in this study will use Hedges' g instead of the traditional Cohen's d.

Primary Analysis

The objectives of the study are to investigate the effectiveness of online peer tutoring groups sizes 1:2 and 1:4 compared to the active control 1:1. A linear mixed effects model fit by REML (Restricted Maximum Likelihood) will be used for the primary analysis. The model used pre- and post-assessments as fixed and schools as random effects.

Secondary analysis

The effect sizes from each of the cohorts will be aggregated in a prospective cumulative meta-analysis.

Process evaluation

The process evaluation will have three purposes. Firstly, it will assess the fidelity of the delivery of the intervention, as this is fundamental prerequisite for the intervention. Secondly, it addressed the feasibility of the intervention for an aggregated trials methodology. Thirdly, it will support the impact evaluation to determine if the group sizes 1:2 and 1:4 can be as effective as 1:1 online peer tuition.

The process evaluation questions are:

- How feasible is it to implement an online cross-age peer tuition project across the transition boundary of primary to secondary school as an aggregated trial?
- How feasible and acceptable do teachers feel it is for using online peer tuition as a transition intervention?
- What are the barriers to the successful implementation of online cross-age peer tuition as a transition intervention? How can these be minimised in future aggregated trials?
- What are the views of the intervention, of the peer tutors, tutees, teachers and IT technicians?
- What are the staff perceptions of the current and possibly sustained impact of the intervention on children's educational attainment?
- What are their suggestions for change if the intervention was to be more widely implemented?

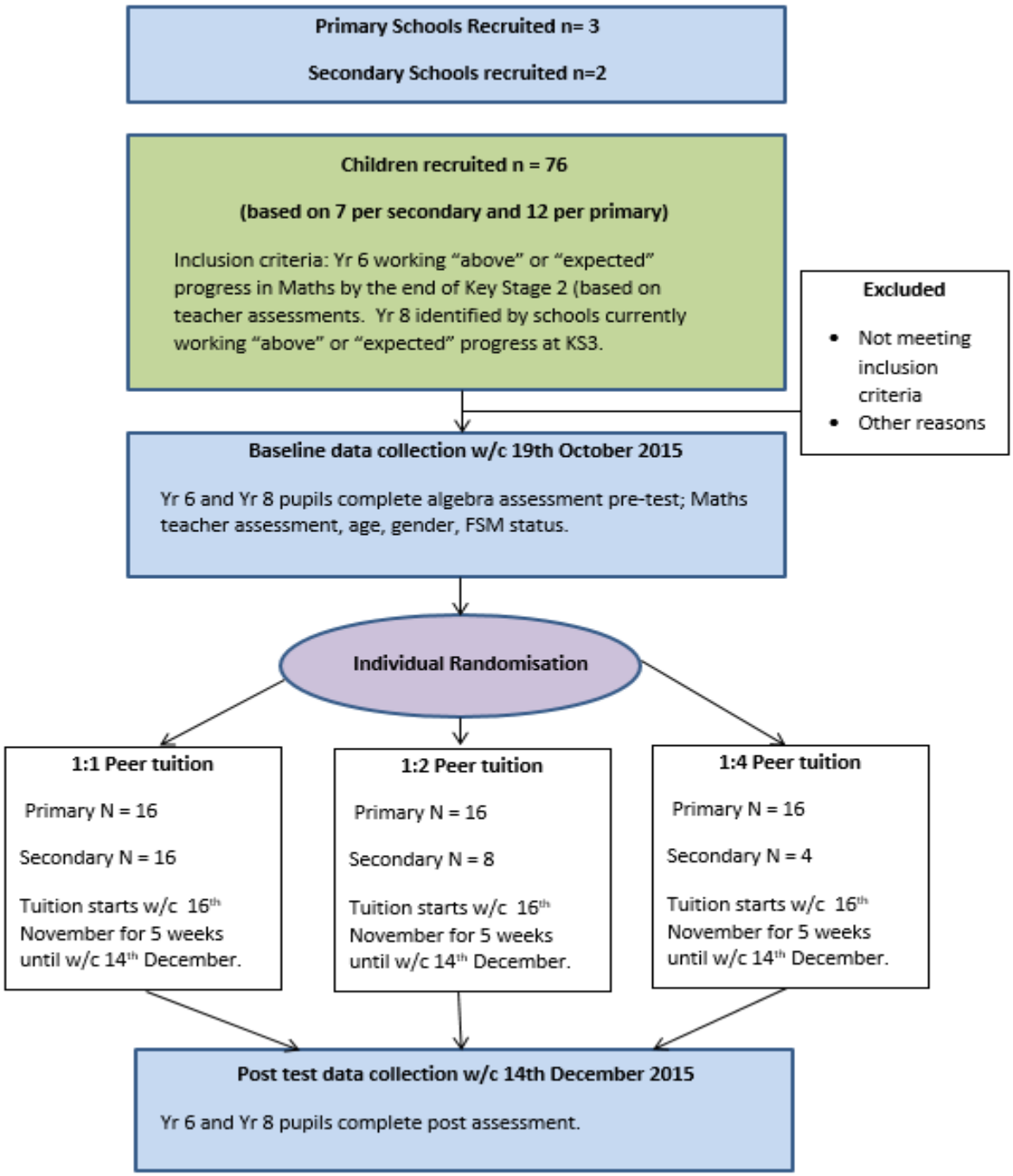
Risks

The main risks associated with this project include:

1) School and pupil recruitment – as a PhD study and with a limited research team of a single researcher, recruiting schools and pupils for the research is a risk.
2) IT – As the intervention is online, the research requires access to the online platform to conduct the research and schools require internet, firewalls to be cleared and suitable pc's.
3) Resources – Funding is a limitation as this is a PhD study, therefore pragmatic decisions will be required in the coordination of the trial and process evaluation.

Appendices

Appendix A: Trial diagram



Appendix B: Trial timelines matrix

Sep – Oct 15	Nov – Dec 15	Jan – Feb 16	Mar – Apr 16	May – Jun 16	Jun – July 16
Ethics approval	Cohort 1 delivery of online peer tutoring	Analysis of data	Cohort 2 delivery of online peer tutoring	Analysis of data	Cohort 3 delivery of online peer tutoring
School recruitment for cohort 1	School visits	School recruitment for cohort 2	School visits	School recruitment for cohort 3	School visits
Resource creation	Post-test	IT testing for recruited schools	Post-test	IT testing for recruited schools	Post-test
IT testing for recruited schools	Interviews and focus groups	Pre-test and randomisation	Interviews and focus groups	Pre-test and randomisation	Interviews and focus groups
Pre-test and randomisation	Ethics approval to be sought for next trials with recording of data.		Ethics approval to be sought for next trials with recording of data.		

Appendix C: School information

Primary School Agreement to Participate

- ☐ I confirm I have read and understood the information sheet for the above project and have had the opportunity to ask questions;
- ☐ I understand that all children's results will be kept confidential and protected using encryption software and that no material which could identify individual children or the school will be used in any reports of this project;
- ☐ I agree for the pupils to undertake an assessment pre and post completion of the project;
- ☐ I agree for the pupils to complete assessments under exam conditions at agreed dates / time;
- ☐ I agree to providing the information letter to all parents of children involved in the project;
- ☐ I consent to Wayne Harrison from Durham University having access to the pupil data to be used solely to evaluate the impact of the project;
- ☐ **I consent to the school taking part in the above study**

Name of head teacher

……

Name of School

……

School Tel Number

……..

Name of school contact

……

School contact Email address

……

Signature of Head teacher

…… Date……..

Thank you for agreeing to take part in this research. Please return this consent form to Wayne Harrison (Durham University).

Secondary School Agreement to Participate

- ☐ I confirm I have read and understood the information sheet for the above project and have had the opportunity to ask questions;
- ☐ I understand that all children's results will be kept confidential and protected using encryption software and that no material which could identify individual children or the school will be used in any reports of this project;
- ☐ I agree for the pupils to undertake an assessment pre and post completion of the project;
- ☐ I agree for the pupils to complete assessments under exam conditions at agreed dates / time;
- ☐ I agree to providing the information letter to all parents of children involved in the project;
- ☐ I consent to Wayne Harrison from Durham University having access to the pupil data to be used solely to evaluate the impact of the project;
- ☐ **I consent to the school taking part in the above study**

Name of head teacher

……

Name of School

……

School Tel Number

……..

Name of school contact

……

School contact Email address

……

Signature of Head teacher

…… Date……..

Thank you for agreeing to take part in this research. Please return this consent form to Wayne Harrison (Durham University).

Dear Parent / Carer

Your child's school is taking part in the *Online Maths Peer Tuition* pilot study as part of a Durham University PhD run by Wayne Harrison, under the supervision of Prof Steve Higgins (s.e.higgins@durham.ac.uk). The aim of the pilot study is to investigate if peer tuition group size has an effect on attainment in online Mathematics tuition. The study will help inform a future trials looking at online peer to peer interventions for schools. The tuition will be delivered using the Tute learning platform, consist of 5 x 30 minute lessons and will last for 5 weeks. If you would like to find out more about the platform please visit www.tute.com.

The *Online Maths Peer Tuition* programme is designed to improve children's maths skills, allowing more able and talented (MAT) Year 6 students to receive tuition from MAT Year 8 pupils. Good maths skills are important for all children. To find out if group size has an effect on *Online Maths Peer Tuition* pupils will be randomly selected to participate in either a one to one, one to two or one to four online peer tuition programme delivered by Year 8 pupils from your local secondary school. In order to test the impact of the type of tuition I will compare the results of the pre and post-tests. In order to complete the evaluation I would like to collect information about your child from your child's primary school. Your child's school will provide information including your child's name, date of birth, gender, unique pupil number, details on your child's current National Curriculum maths level and free school meal status.

Your child's information will be treated with the strictest confidence. I will not use your child's name or the name of the school in any report arising from the research. Your child's information will be kept confidential at all times. This study has been reviewed and approved by the School of Education Ethics Sub-Committee at Durham University on 25/9/2015.

If you are happy for your child's information to be used you do not need to do anything. Thank you for your help with this project.

If you would rather your child's school did not share your child's information for this project please complete the enclosed opt out form and return it to your child's school by 22.10.2015.
If you would like further information about the *Online Maths Peer Tuition* project please contact Wayne Harrison, NEDTC Doctoral Education Student: w.d.harrison@durham.ac.uk, 0191 3342000.

Yours faithfully

W. D. Harrison

W D Harrison (NEDTC Doctoral Student Durham University)

Professor Steve Higgins (Durham University)

TUTE

Online Maths Peer Tuition Pilot Study: Opt Out Form

If you **DO NOT** want your child's data to be shared for use in the Online Maths Peer Tuition study, please return this form to your child's school asap.

☐ **DO NOT** want my child's data to be shared for use in the Online Maths Tuition study.

Parent/Carer Signature: ……

Date: ……..

Child's Name: ……

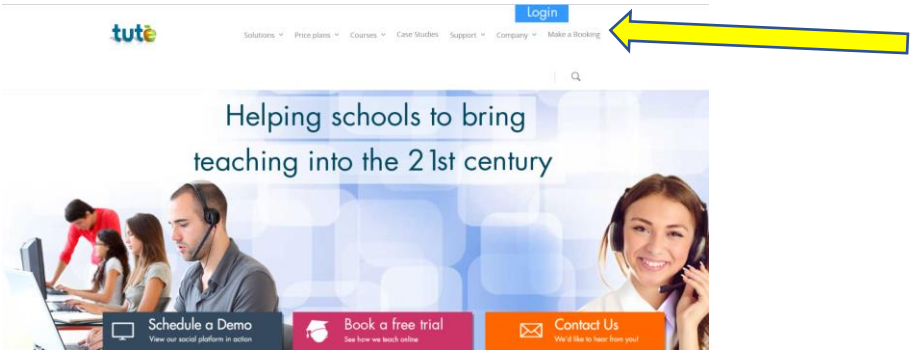
Child's School: ……..

Any concerns about this study should be addressed to the Ethics Sub-Committee of the School of Education, Durham University via email (Prof Steve Higgins, School of Education, tel. (0191) 334 8403, e-mail: s.e.higgins@Durham.ac.uk).

School guidance document for primary teachers

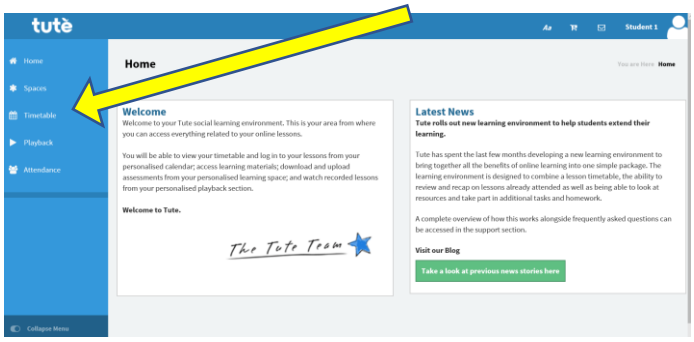
Peer tuition – Primary school guide to starting a session

1) Turn on your computer and plug in your headsets (make sure you place these into the computer correctly, the sound is turned on and not muted)
2) Open up an internet browser (Google, IE or Firefox) and type in www.tute.com
3) Click on the log in button as shown below:

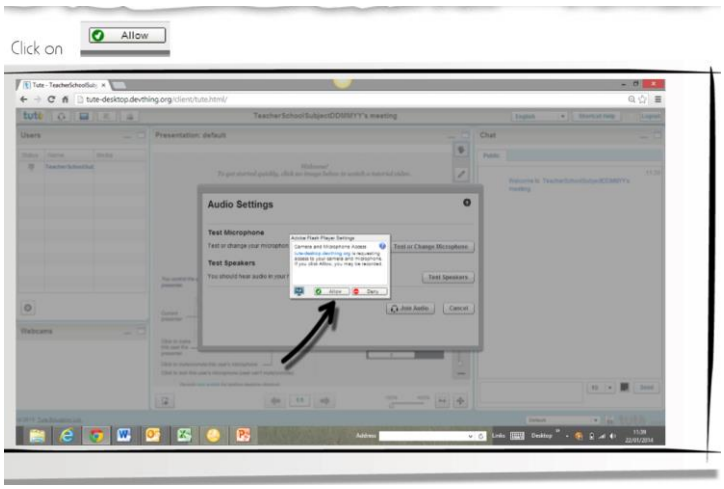


Click here and type in your username and password (case sensitive)

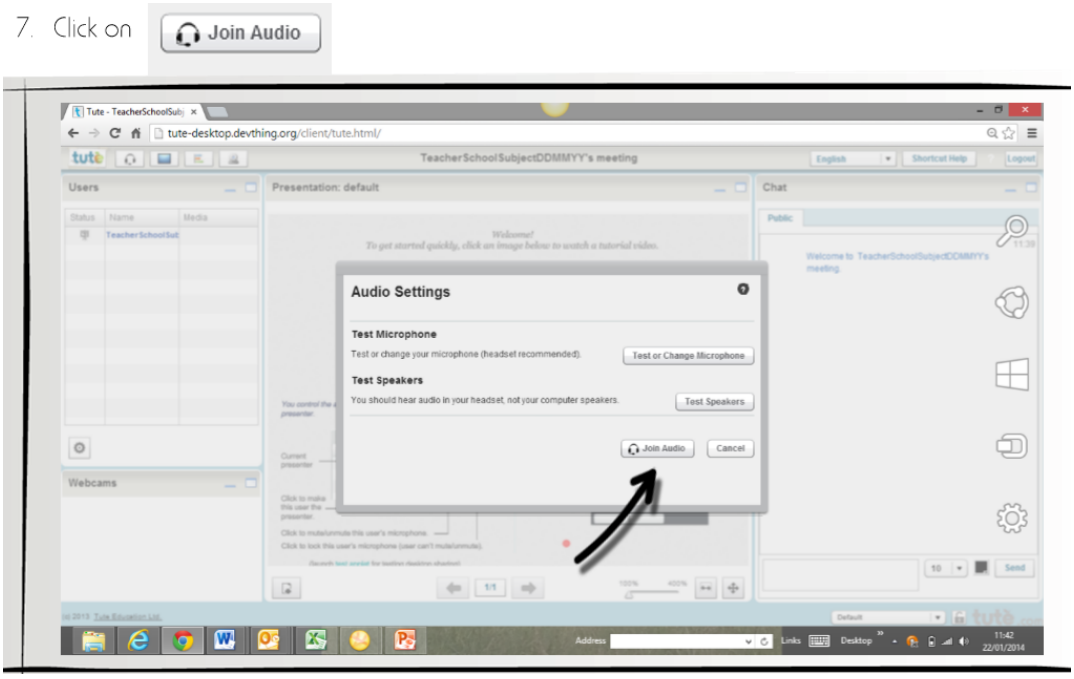
4) Click on the timetable on the left hand side.



5) Click on the link that is on the time table screen for the day and time of your lesson.
6) The Tute virtual classroom will now open then click the allow button.



7) Click test the microphone and then test the speakers. Then click join audio.



8) Enjoy your lesson ☺
9) If you cannot hear your teacher please try the following:

Online Cross-age Peer tuition assessment

Name: _____

Date: _____

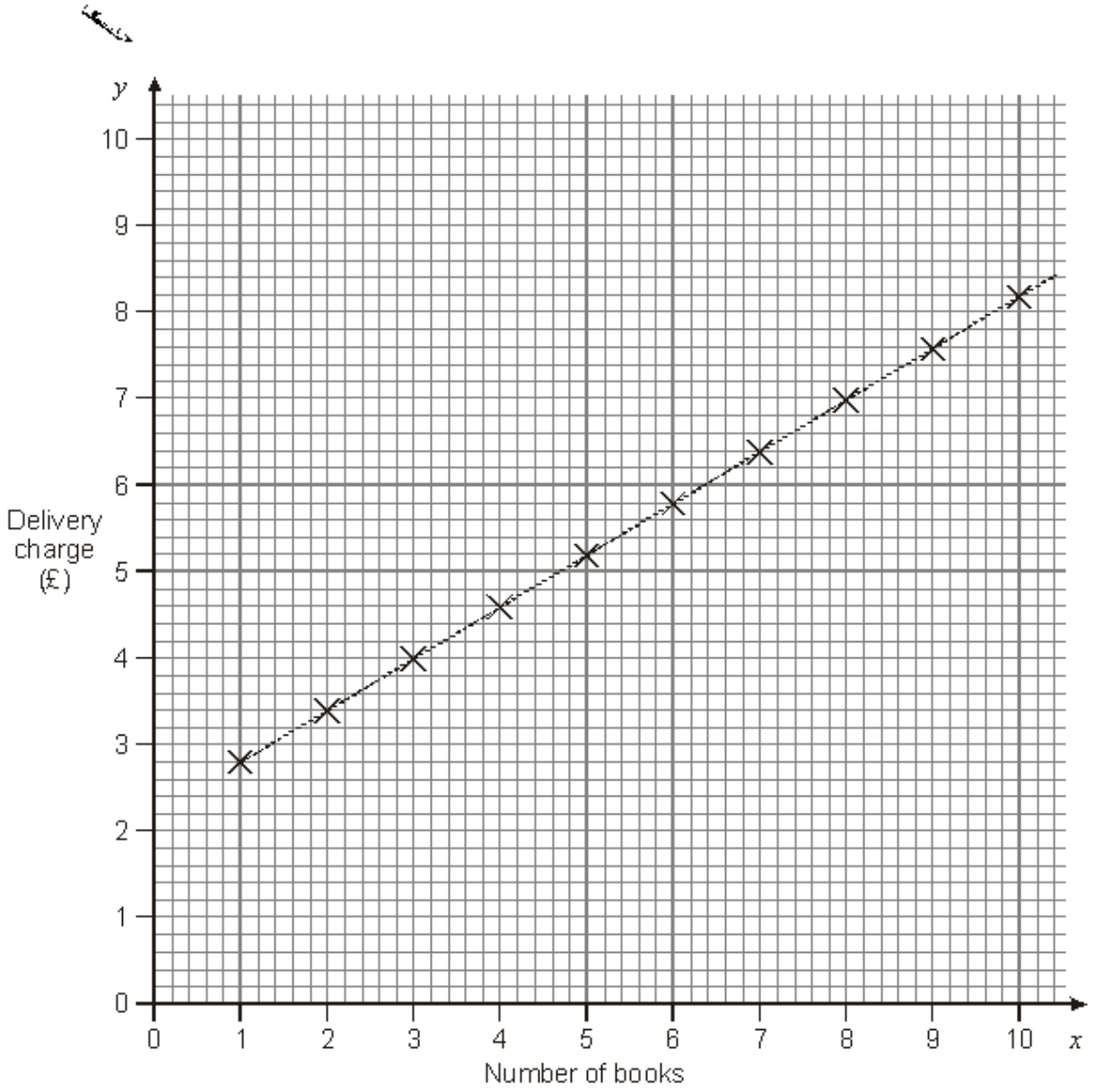
Year group: _____

School: _____

Question 1

A company sells books using the internet.

The graph shows their delivery charges.



(a) Use the graph to fill in the values in this table.

Number of books	Delivery charge (£)
8	
9	

1 mark

(b) For every extra book you buy, how much more must you pay for delivery?

..... p

1 mark

(c) A second company sells books using the internet.

Its delivery charge is **£1.00 per book**.

On the graph opposite, draw a line to show this information.

1 mark

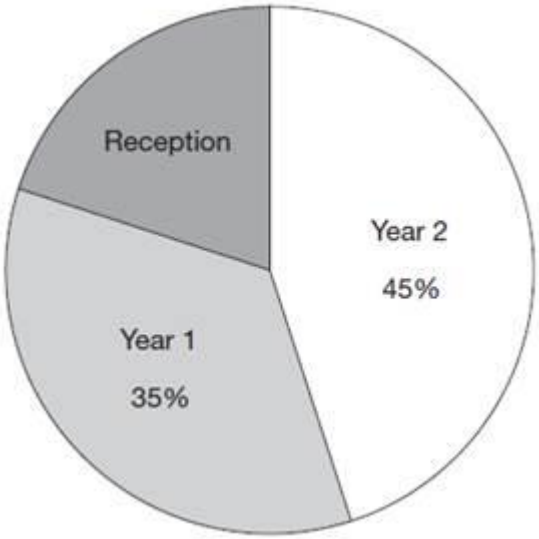
(d) Complete the sentence.

Delivery is cheaper with the **first** company

if you buy at least ..... books. 1 mark

Question 2

The pie chart shows the Year groups of children at Woodland Infant School.



There are **56** children in **Year 1**.

How many children are there in Reception?

2 marks

Question 3

Paul is making a **pie chart** of land use in Great Britain using these survey results.

Land use	Area in thousands of square kilometres
Built–up areas (towns and cities)	33
Farms	133
Woodland	68
Open space	6

Calculate the **angle** of the sector for **farms**.

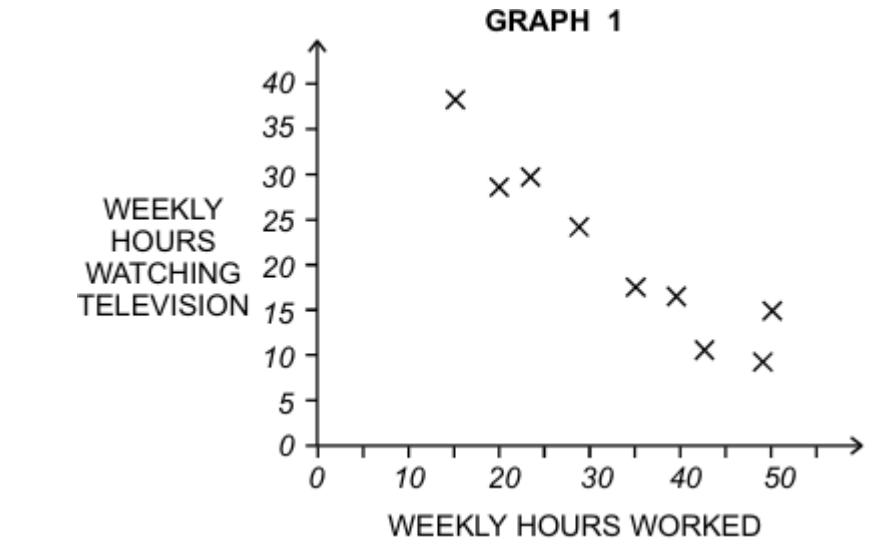
2 marks

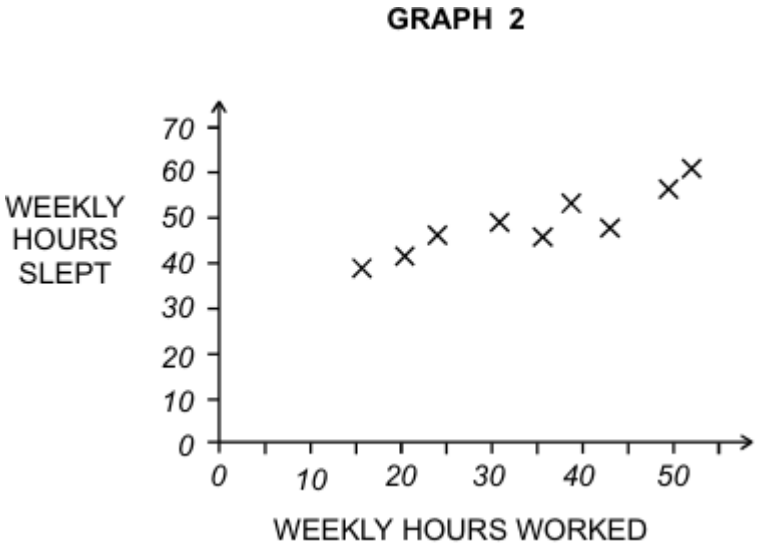
Question 4

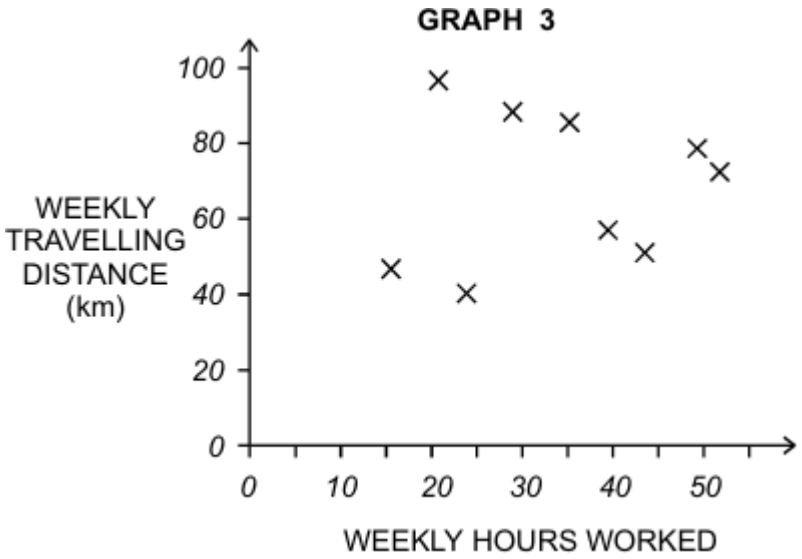
Nine students were discussing their holiday jobs working on a local farm.

They decided to find out if there were any relationships between the time they spent working, sleeping, watching television and the distance they had to travel to work.

The students plotted three scatter graphs.







(a) What does **graph 1** show about the relationship between the weekly hours spent watching television and the weekly hours worked?

1 mark

(b) What does **graph 2** show about the relationship between the weekly hours slept and the weekly hours worked?

1 mark

(c) What does **graph 3** show about the relationship between the weekly travelling distance and the weekly hours worked?

1 mark

(d) Another student works 30 hours per week.

Use **graph 1** to estimate the weekly hours spent watching television by this student.

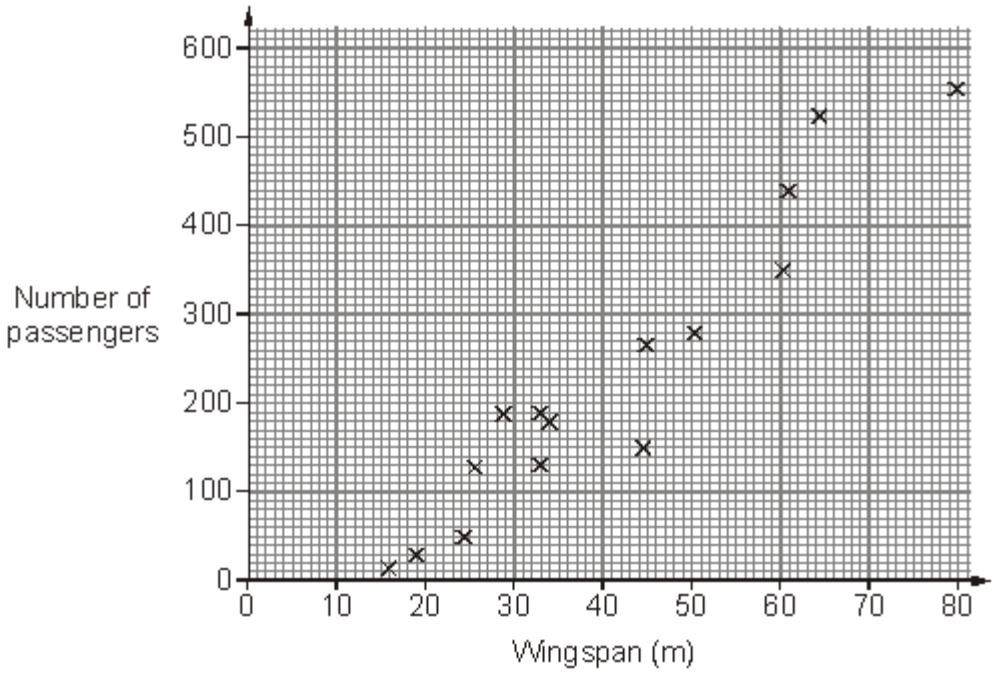
Explain how you decided on your estimate.

2 marks

Question 5

Planes

The scatter graph shows the maximum number of passengers plotted against the wingspans of some passenger planes.



(a) What type of correlation does the scatter graph show?

.....

1 mark

(b) Draw a **line of best fit** on the scatter graph.

1 mark

(c) Another passenger plane has a **wingspan** of **40 m**. The plane is full of passengers.

If each passenger takes **20 kg** of bags onto the plane, estimate how much their
bags would weigh altogether.

..... kg

2 marks

Data handling assessment - Mark scheme

Q1

(a) Completes the table correctly, i.e.

8 7.(00)

9 7.60

Accept for 9 books, a value between 7.55 and 7.65 inclusive

! 7.60 shown as 7.6

Condone

1

(b) 60 p

! Follow through from part (a)

Accept provided their 7.60 > their 7.00

1

(c) Draws the correct straight line $y = x$, at least of length 6cm, including the point of intersection with the given line, with no errors

! Line not dashed

Condone

! Line not ruled or accurate

Accept provided the pupil's intention is clear

Do not accept series of points that are not joined

1

(d) 6

! Follow through from an incorrect line in part (c)

Provided there is only one point of intersection, follow through as the closest integer value above their *x*-value
e.g., from their intersection as (7.2, 6.5), accept

- *8*

e.g., from their intersection as (4, 4.6), accept

- *5*

! Maximum of 10 books assumed

Condone
e.g., accept

- *6 to 10 books*

U1

[4]

Question 2

32

2

or

160 seen *(the total children in the school)*

Do not accept 160° or 160%

OR

Shows or implies a complete, correct method, e.g.:

- 35 + 45 = 90 *(error)*
  100 − 90 = 10
  56 ÷ 35 = 1.6
  1.6 × 10 = 16

- 35% of children = 56
  total children = 56 × 100 ÷ 35 = 150 *(error)*
  Reception = 100 – (45 + 35)% = 20%
  Reception = 20% of 150
  0.2 × 150 = 40 *(error)*

- 35% is 56
  5% is 8
  20% is 4 × 8 = 24 *(error)*

1

[2]

Question 3

Award **TWO** marks for the correct answer of 199.5

*Accept 199 **OR** 200° **OR** unrounded values, e.g. 199.499*

If the answer is incorrect award **ONE** mark for evidence of an appropriate method, e.g.

- 33 + 133 + 68 + 6 = 240 **AND** 360 ÷ 240 × 133.

The calculation need not be completed for the award of the mark.

up to 2

[2]

Question 4

(a) States that the relationship is a negative one, e.g.:

- The more hours worked, the fewer hours spent watching television.
- The less hours worked, the more hours spent watching television.

- As television goes down, weekly hours goes up.
- As one goes down the other goes up.

Accept negative relationship or negative (inverse) correlation.

A description of the graph e.g.

- *It goes down*

Do not accept responses that do not include reference to changes in both variables e.g.

- *Watching TV is bad for your work.*
- *You can't watch television and work.*

1

(b) States that the relationship is a positive one, e.g.:

- As hours worked increases, the hours slept increases slightly.
- The more hours slept, the more hours worked.
- More work, more sleep.
- As one goes up so does the other

Accept positive relationship or weak positive relationship
or positive (direct) correlation.

A description of the graph e.g.

- *It goes up slowly.*

Do not accept responses that do not include reference to changes in both variables e.g.

- *Sleeping a lot is good for your work.*

1

(c) States that there is no apparent relationship, e.g.:

- The travelling distance is not related to the hours worked.
- There is no connection.
- It does not matter.

Accept no relationship or none or no correlation.

Do not accept a description of the graph e.g.:

- *The points are all over the place.*

Do not accept explanations which imply that the graph shows no information about the relationship e.g.:

- *Nothing*

1

(d) **1m** states any value between 20 and 25 inclusive.

1m explains that a value is found by considering the crosses to one or both sides of the 30 hours mark and indicating a point that fits in; or

that a line is drawn through the crosses, then a line is drawn up from 30 hours to meet this line and the value read across from that point; or that the general pattern of the graph was considered, e.g.:

- I put it between the crosses either side of the mark.
- Because there is another result close to the 30 hours.
- I went to 30 and up to the line and across to the line.
- 25 was on my best fitted line.
- I put a cross in where it fitted in with the rest of the line.
- It follows the pattern of the graph.

It is not necessary for reference to be made to 30 hours or values of any of the crosses shown to be given.

Accept responses where a written explanation is not given but where the graph is annotated, either by a point being made between the crosses either side of the 30 hours mark or by a line of best fit being drawn through the points.

It is not necessary for the line of best to fit to be straight.

2

[5]

Question 5

(a) Indicates the correlation is positive

! Positive qualified

Ignore

e.g., accept

- *Strong positive*
- *Direct positive*

Do not accept sign of correlation not indicated

e.g.

- *High*
- *Strong*

! Relationship quantified

Ignore alongside a correct response

Do not accept relationship described without reference to correlation

e.g.

- *The greater the wingspan, the more passengers it can hold*

1

(b) Draws a line of best fit within the tolerance, and at least of the length, as shown on the overlay

! Line not ruled or accurate

Accept provided the line is within tolerance, and at least of the length required

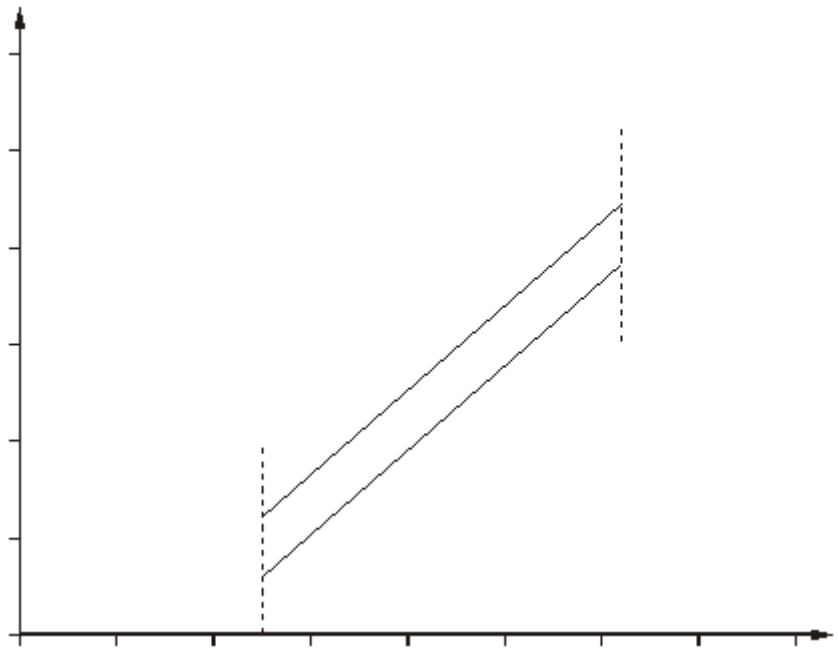
! Line of best fit is incorrect beyond the dashed lines on the overlay

Condone

e.g., accept

- *A correct line of best fi t that is then joined to the origin*

Overlay



1

(c) 3600 to 5200 inclusive

2

or Shows a value between 180 and 260 inclusive

or

Shows a value that follows through from their line of best fit

e.g.

- Their line passes through the point (40, 280), final answer: 5600

! For 1m, range for follow-through value

If their line goes through (40, y) accept follow-through as 20 × (y ± 10) provided their line always has a positive gradient

1

(U1)

[4]

Appendix 3: Statistical Analysis Plan Study One

Pilot efficacy study for small scale aggregated trials investigating the effectiveness of online cross-age peer tuition group sizes across the transition boundary between primary and secondary schools (Study 1).

STATISTICAL ANALYSIS PLAN

Version 1.1

Version date: December 2015

Author: Wayne Harrison

Durham University

Note: This analysis plan was written post-randomisation and prior to collection of any outcome data.

Definition of terms

ITT: Intention to treat

Tute: Online education company

Trial Objectives

The trial aims to determine the feasibility of implementing a more able and talented cross-age peer tuition intervention across the transition boundary between primary and secondary schools.

Primary Objective

The primary research questions is:

2) Is the implementation of an online cross-age peer tuition intervention feasible between primary and secondary schools as an aggregated trial?
    d) Is online cross-age peer tuition a feasible intervention strategy between schools as part of an aggregated trial?
    e) Are the instruments used appropriate for the ability of students involved in the study?
    f) Can the intervention maintain fidelity if delivered to schools in other locations in the UK?

Secondary Objectives

Secondary objectives are:

3) Does one-to-two online cross-age peer-tutoring produce equivalent effects in mathematics attainment[26] compared with online one-to-one tuition?
4) Does one-to-many online cross-age peer-tutoring produce equivalent effects in mathematics attainment[2] as compared with one-to-one tuition?

[26] Mathematics attainment is defined as mastery level concepts in KS2 curriculum for Algebra.

Design

The design for the study will be a pragmatic randomised controlled trial using an aggregation of three smaller trials over a sequential timescale of 9 months. The individual trials will be conducted as an efficacy study as the intervention is required to be conducted under highly controlled and optimal conditions in order to maximise outcomes, and ensure the technology worked in schools.

Full details of the background and trial design can be found within the protocol (Harrison, 2016).

Sample size

Prior to randomisation, each participating secondary school will identify 7 pupils and each primary school identified 12 pupils who are eligible to participate in the trial. Since the removal of the national framework of 'levels' to report pupils attainment and progress it was not possible to specify specific inclusion criteria. A relative criterion, depending on each individual school assessment – "above expected", "expected" and "below expected" progress will be used. Consequently, consistency between individual schools is a potential limitation.

The inclusion criteria used in the study are:

School inclusion criteria: Secondary and primary schools are eligible to take part in the trial if they agree to all trial procedures including provision of pupil data, informing parents, randomisation and implementation of the intervention.

Primary pupil inclusion criteria: Pupils eligible for the intervention will be either at "expected" or "above expected" for progress in Mathematics by the end of Year 6, based on teacher assessments.

Secondary pupil inclusion criteria: Pupils eligible for the intervention will be either at "expected" or "above expected" for progress in Mathematics by the end of Year 8, based on teacher assessments.

Randomisation

Randomisation will be performed at an individual level in each of the participating schools and carried out by an independent person[27]. The reason randomisation is blinded is to prevent selection bias, as

[27] A qualified teacher not involved in the study performed the randomisation, after performing a trial run under the supervision of the author of this study to ensure the process was completed correctly.

a recent systematic review found that trials using non-blinded assessors generated more optimistic effect sizes than blinded assessors (Hróbjartsson et al., 2014).

Unequal randomisation will be used for the peer tutors in each secondary school participating in the study, with 1 pupil assigned to the 1:4 group, 2 pupils assigned to the 1:2 group and 4 pupils assigned to the 1:1 group. For each participating primary school, the pupils will be ranked in order of performance based on the data handling pre-assessment as a measure of mathematics ability for this specific topic, with 1 equating to the highest and 12 the lowest scores. These will be divided into groups of three, starting with 1, 2, and 3. These will then be randomly selected by the independent person to be allocated into the group sizes 1:1, 1:2 and 1:4. This process will be repeated for blocks 4, 5, 6 and 7, 8, 9, then finally 10, 11, 12. Through blocking the performance ability as indicated by the pre-assessment on data handling this should result in an equal distribution of ability across the groups.

Outcomes

The assessment of the primary outcome used will be designed by the researcher in partnership with a Year 6 teacher from school B, using previous standardised SAT assessments from the schools TestBase assessment system. The assessment will cover data handling in mathematics and has a maximum mark of 17.

Data

Baseline data

Pupil baseline data for the intervention and control groups will be presented by trial arm to assess balance. Pupil level characteristics will be summarised by trial arm for both, as randomised, and included in the primary analysis. Information on the lessons from which the pupils were withdrawn is also summarised. No formal statistical testing to assess balance was conducted.

Outcome data

The mathematics assessment will be delivered on an agreed date prior to the start of the intervention and in the final week of the planned online lesson. The assessments were completed under ‘exam conditions’ with all pupils (control and intervention) participating at the same time.

Cleaning and formatting

Baseline data will be checked upon receipt for completeness and to ensure pupils are eligible to take part in the trial.

Analysis

Analysis will use intention to treat, therefore all pupils will be analysed in the group they were randomly allocated regardless of whether or not they received the intervention. Analyses will be conducted using R statistical software, using a package developed by Durham University which is commonly used in EEF evaluations. Effect sizes are presented relating to the analyses alongside 95% confidence intervals. The effect size is defined as:

$$g = \frac{\bar{x}_1 - \bar{x}_2}{s^*}$$

where the pooled standard deviation *s** is computed as:

$$s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

Trial Completion (CONSORT flow diagram)

A CONSORT diagram will be produced to show the flow of schools and pupils through the trial. This will include the number of pupils opting out of the trial.

Primary analysis

A linear mixed effects model fit by REML (Restricted Maximum Likelihood) will be used for the primary analysis. The model used pre- and post- assessments as fixed and schools as random effects. The justification for schools as random is because it is assumed the effect varies randomly within the population of the organisation. The reasoning for selecting pre- and post-assessments as fixed is because it is assumed to be measured without measurement error and the variable used contains most or all of the variable values in the population.

Secondary analyses

Cumulative meta-analysis

Two additional cohorts in Spring 2016 and Summer 2016. The data will be combined with the feasibility study to model the development of a prospective cumulative meta-analysis using a number of models.

Compliance

Non-compliance will be summarised in terms of student attendance (number of lessons attended) with thresholds at 75% and 50%. In addition, the number of pupils attending 75% and 50% of sessions on time will be summarised.

Appendix 4: Study One Exploratory Analysis of Group Size

The pre-test and post-test data from the two schools who were unable to participate in the intervention due to technical issues were used to create a retrospective control to explore the impact of different group sizes of online peer tutoring against a business as usual control. The group sizes 1:1, 1:2 and 1:4 were compared to the business as usual group.

Appendix 10 includes the raw data used in the analysis using a linear mixed effects model fit by REML (Restricted Maximum Likelihood). This analysis is completed outside of the SAP and is excluded from the main thesis, but contains potentially useful information in relation to cost/benefit analysis.

	1:1	1:2	1:4
Effect Size (Hedges' g)	1.44 (0.03 to 2.86)	1.02 (0.1 to 1.95)	1.16 (-0.24 to 2.56)
School Standard Deviation Intercept (residual)	2.21 (2.41)	0.88 (2.41)	2.22 (2.52)
Intra-class correlation	0.46	0.12	0.44
Pre-test			
Value	0.35	0.77	0.54
Standard error	0.11	0.12	0.15
DF	38	29	28
t-value	3.11	6.40	3.58
p-value	<0.00	0.04	<0.00
Post-test			
Value	0.35	2.69	4.33
Standard error	0.11	0.58	2.25
DF	3	3	3
t-value	2.43	4.66	1.95
p-value	0.09	0.02	0.15
Total sample size	44	35	34

The effect sizes of the groups 1:1, 1:2 and 1:4 compared to the business as usual are shown in the table above. The effect of group size 1:1 shows a difference of 1.44 (95% CI: 0.03 to 2.86) in the mathematics performance as defined by the data handling assessment, whereas the group size 1:2 is 1.02 (0.1 to 0.1.95). The 1:4 groups size shows an effect size of 1.16 (-0.24 to 2.56) compared to the business as usual group. All three demonstrate a large effect size but these should be treated with some caution. Possible reasons for the large effect size could be due to the high level of support in ensuring the intervention was delivered correctly with high fidelity, the intervention was delivered over a very short timescale with a highly focused intervention and the comparator group did not receive any form of interventional support. However, these are tentative indications of the benefit of peer tutoring in larger groups (one tutor with two or event four students) and further studies should be conducted comparing different group sizes for online peer tutoring with matched business as usual control group. This potentially changes the cost/benefit of the approach, reducing any tutoring costs and increasing the number of student who can receive tutoring.

Protocol

A feasibility study to develop a prospective cumulative meta-analysis based on aggregated small scale randomised controlled trials to evaluate the effectiveness of a Tute mathematics intervention

Version: 1.0

Date: May 2016

Study lead: Wayne Harrison (Durham University)

Contact: w.d.harrison@durham.ac.uk

Background and significance

The study is funded as part of an ESRC funded PhD at Durham University to investigate the effectiveness of online small group teaching in mathematics. The study is developing a new methodology for aggregating scale small trials into a prospective cumulative meta-analysis.

The intervention will provide pupils in Year 7 online small group tuition in the subject of mathematics, using qualified teachers from TUTE to deliver the lessons. The online peer tuition lessons will be delivered weekly in 45 minute lessons via an online platform.

The intervention aims to help improve attainment in mathematics for Year 7 pupils.

Research question

The trial aims to determine the effectiveness of online small group teaching in mathematics.

Primary Objective

The primary research questions is:

What is the effectiveness of a small group online mathematics programmes compared with "business as usual" on the maths skills of participating children?

Secondary Objectives

Secondary objectives are:

4) Does dosage impact on the effect of the intervention?
5) Is there a differential impact on pupil premium students[28]?
6) Is it feasible to implement a prospective cumulative meta-analysis for the aggregation of small scale randomised controlled trials in schools?

[28] Pupil Premium children are identified as coming from disadvantaged backgrounds, such as low income families, looked after care or children from service families.

Design

The design for the study will be a pragmatic randomised controlled trial using an aggregation of three smaller trials over a sequential timescale of 9 months. The individual trials will be conducted as an efficacy study as the intervention will be required to be conducted under highly controlled and optimal conditions in order to maximise outcomes, and ensure the technology worked in schools.

This will be a waiting list design. This design is ethical as it does not deprive anyone of the treatment. It is less likely to demoralise schools / pupils in the control group than had there not been a waiting list. If the intervention is successful, the pupils will receive the intervention in the next academic year.

Recruitment

The schools initially targeted for participation in this trial will be based in the North East of England, due to logistical reasons as this trial ran simultaneously with cohort 3 of the primary study 1. The research lead will provide an information document for parents and pupils and it is anticipated that the intervention will be delivered in Summer term 2, Autumn term and Spring terms for the three cohorts of schools.

Inclusion criteria

The following inclusion criteria used in the study are:

School inclusion criteria: Secondary schools are eligible to take part in the trial if they agree to all trial procedures including provision of pupil data, informing parents, randomisation and implementation of the intervention[29].

Pupil inclusion criteria: Pupils eligible for the intervention will be either "below expected" or "expected" for progress in Mathematics by the end of Year 7, based on teacher assessments.

Each school will provide parents with an information letter and an optional opt out form if they do not wish their child to participate in the research. Please see the appendix for examples of these documents.

[29] Appendix 5 includes the 'Agreement to Participate Form'

School participation

Schools are required to sign a memorandum of understanding which outlines the following:

- I confirm I have read and understood the information sheet for the above project and have had the opportunity to ask questions;
- I understand that all children's results will be kept confidential and protected using encryption software and that no material which could identify individual children or the school will be used in any reports of this project;
- I agree for the pupils to undertake an assessment pre and post completion of the project;
- I agree for the pupils to complete assessments under exam conditions at agreed dates / time;
- I agree to providing the information letter to all parents of children involved in the project;
- I consent to Wayne Harrison from Durham University having access to the pupil data to be used solely to evaluate the impact of the project;

Intervention

The intervention is a mathematics programme designed by TUTE to deliver online small group lessons in virtual classrooms.

Randomisation

Randomisation will be conducted at pupil level in each school, carried out by an independent individual (independent teacher). The 24 pupils will be ranked in order of performance using Alfiesoft mathematics assessment pre-test (1 equating to best and 24 the least) and paired on ability. A coin will be used to select one of each successive pair of classes starting with 1&2 (then 3&4, etc.) forming the control and intervention group. Through blocking the performance ability as indicated by the Alfiesoft pre-assessment on mathematics performance this will result in an equal distribution across the groups.

Sample size calculation

The focus of this trial is on pupils who are performing at "below" or "expected" progress, therefore the sample size was based on this subgroup of children and limited to 12 pupil places in the TUTE

virtual classroom. The initial cohort for the cumulative meta-analysis involves 5 schools, with 24 pupils selected from Year 7. The total sample will be 120 pupils, 60 receiving the intervention and 60 continuing as business as usual. A recent EEF evaluation (Torgerson et al., 2014) found a positive effect size for small group tuition (weighted mean effect size[30] of 0.34). The aggregated study is powered to have an 80% chance of detecting an effect size of 0.34, with a sample of 360 pupils. However, through the prospective nature of a cumulative meta-analysis, repeating the trial and aggregating the results would allow for a more robust analysis of effectiveness. The intra-class correlation rho = 0.15 accounts for the variance between schools, rho = 0.15 is selected as this is the recommended intra-class correlation for achievement in Mathematics (Tymms, 2012).

Outcome measures

The primary outcome measure will be the mathematics attainment scores from the Alfiesoft assessment and this will be marked blind by the online platform.

Analysis

Impacts will be estimated on the basis of intention to treat, whereby all pupils who are initially involved in the pre – and post-testing will be analysed according to the group they were initially assigned, regardless of whether they went on to participate in the intervention.

Effect sizes will be calculated and presented alongside their 95% confidence intervals. The effect size reported in this study will use Hedges' g instead of the traditional Cohen's d.

Primary Analysis

The primary objective of this study will be to investigate the effectiveness of the TUTE intervention on the mathematics skills of the pupils in the intervention group. The primary measure will be the Alfiesoft mathematics scores and this was analysed using a multi-level model.

[30] Effect size used Cohen's d.

Secondary analysis

A secondary analysis using the data from the intervention group using a multi-level analysis will be completed to determine the impact of the intervention on pupil premium funded students. An analysis on the dosage and behaviour for learning will be included.

The effect sizes from each of the cohorts will be aggregated in a prospective cumulative meta-analysis.

Process evaluation

The process evaluation will contain two main purposes. Firstly, it will assess the fidelity of the delivery of the TUTE online intervention, as this is a fundamental prerequisite for the intervention. Secondly, it will address the feasibility of conducting an aggregated trial for the effectiveness of an online small group tuition intervention.

The process evaluation questions are:

- How feasible is it to evaluate the effectiveness of an online intervention using an aggregated trial methodology?
- How feasible and acceptable do teachers feel it is for using online small group tuition as a numeracy catch-up intervention?
- What are the barriers to the successful implementation of online small group tuition? How can these be minimised in future aggregated trials?
- What are the views of the intervention, of the peer tutors, tutees and teachers?
- What are the staff perceptions of the current and possibly sustained impact of the intervention on children's educational attainment?
- What are their suggestions for change if the intervention was to be more widely implemented?

Risks

The main risks associated with this project include:

4) School and pupil recruitment – as a PhD study and with a limited research team of a single researcher, recruiting schools and pupils for the research is a risk.

5) IT – As the intervention is online, the research requires access to the online platform to conduct the research and schools require internet, firewalls to be cleared and suitable pc's.
6) Resources – Funding is a limitation as this is a PhD study, therefore pragmatic decisions will be required in the coordination of the trial and process evaluation.

Appendices

Appendix A: Trial diagram

Appendix B: Trial timelines matrix

May – Jun 16	June – Jul 16	Sept – Oct 16	Nov – Dec 16	Jan – Feb 17	Mar– Apr 17
Ethics approval	Cohort 1 delivery of small group tutoring	Analysis of data	Cohort 2 delivery of small group tutoring	Analysis of data	Cohort 3 delivery of online peer tutoring
School recruitment for cohort 1	School visits	School recruitment for cohort 2	School visits	School recruitment for cohort 3	School visits
Resource creation	Post-test	IT testing for recruited schools	Post-test	IT testing for recruited schools	Post-test
IT testing for recruited schools	Interviews and focus groups	Pre-test and randomisation	Interviews and focus groups	Pre-test and randomisation	Interviews and focus groups
Pre-test and randomisation	Ethics approval to be sought for next trials with recording of data.		Ethics approval to be sought for next trials with recording of data.		

Appendix C: School information

Secondary School Agreement to Participate

- ☐ I confirm I have read and understood the information sheet for the above project and have had the opportunity to ask questions;
- ☐ I understand that all children's results will be kept confidential and protected using encryption software and that no material which could identify individual children or the school will be used in any reports of this project;
- ☐ I agree for the pupils to undertake an assessment pre and post completion of the project;
- ☐ I agree for the pupils to complete assessments under exam conditions at agreed dates / time;
- ☐ I agree to providing the information letter to all parents of children involved in the project;
- ☐ I consent to Wayne Harrison from Durham University having access to the pupil data to be used solely to evaluate the impact of the project;
- ☐ **I consent to the school taking part in the above study**

Name of head teacher

……

Name of School

……

School Tel Number

……..

Name of school contact

……

School contact Email address

……

Signature of Head teacher

…… Date……..

Thank you for agreeing to take part in this research. Please return this consent form to Wayne Harrison (Durham University).

Dear Parent / Carer

Your child's school is taking part in the *Online Maths Small Group* pilot study as part of a Durham University PhD run by Wayne Harrison, under the supervision of Prof Steve Higgins (s.e.higgins@durham.ac.uk). The aim of the pilot study is to investigate the impact of online small group tutoring on attainment in Mathematics. The study will help inform a future trials looking at online maths interventions for schools. The tuition will be delivered using the Tute learning platform, consist of 6 x 45 minute lessons and will last for 6 weeks. If you would like to find out more about the platform please visit www.tute.com.

The *Online Maths Peer Tuition* programme is designed to improve children's maths skills, using qualified teachers from Tute to deliver online intervention lessons. Good maths skills are important for all children. To find out if online small group tuition is effective, pupils will be randomly selected to participate in the online intervention with pupils not selected receiving the intervention in the next academic year. In order to test the impact of the type of tuition I will compare the results of the pre and post-tests. In order to complete the evaluation I would like to collect information about your child from your child's primary school. Your child's school will provide information including your child's name, date of birth, gender, details on your child's current National Curriculum maths level and free school meal status.

Your child's information will be treated with the strictest confidence. I will not use your child's name or the name of the school in any report arising from the research. Your child's information will be kept confidential at all times. This study has been reviewed and approved by the School of Education Ethics Sub-Committee at Durham University on 25/9/2015.

If you are happy for your child's information to be used you do not need to do anything. Thank you for your help with this project.

If you would rather your child's school did not share your child's information for this project please complete the enclosed opt out form and return it to your child's school by 30.5.2016.
If you would like further information about the *Online Maths Small Group* project please contact Wayne Harrison, NEDTC Doctoral Education Student: w.d.harrison@durham.ac.uk, 0191 3342000.

Yours faithfully

W. D. Harrison

W D Harrison (NEDTC Doctoral Student Durham University)

Professor Steve Higgins (Durham University)

TUTE

Online Maths Peer Tuition Pilot Study: Opt Out Form

If you **DO NOT** want your child's data to be shared for use in the Online Maths Small Group study, please return this form to your child's school asap.

☐ **DO NOT** want my child's data to be shared for use in the Online Maths Small Group Study.

Parent/Carer Signature: ……

Date: ……..

Child's Name: ……

Child's School: ……..

Any concerns about this study should be addressed to the Ethics Sub-Committee of the School of Education, Durham University via email (Prof Steve Higgins, School of Education, tel. (0191) 334 8403, e-mail: s.e.higgins@Durham.ac.uk).

Appendix 6: Study Two Statistical Analysis Plan

A feasibility study to develop a prospective cumulative meta-analysis based on aggregated small scale randomised controlled trials to evaluate the effectiveness of a Tute mathematics intervention (Study 2)

STATISTICAL ANALYSIS PLAN

Version 1.1

Version date: May 2016

Author: Wayne Harrison

Durham University

Note: This analysis plan was written post-randomisation and prior to collection of any outcome data.

Definition of terms

ALFIE: Online assessment

CACE:

FSM: Free school meals

ITT: Intention to treat

Tute: Online education company

Trial Objectives

The trial aims to investigate the effectiveness of a TUTE mathematics intervention on the mathematical skills of participating Year 7 pupils struggling with maths compared with Year 7 pupils continuing with 'business as usual' lessons.

Primary Objective

The primary research questions is:

1) What is the effectiveness of small group online mathematics programmes compared with "business as usual" on the maths skills of participating children?

Secondary Objectives

Secondary objectives are:

1) Does the dosage analysis impact on the effect of the intervention?
2) Is there a differential impact on pupil premium students?
3) Is it feasible to implement a prospective cumulative meta-analysis for the aggregation of small scale randomised controlled trials in schools?

Design

This trial will be a pragmatic randomised controlled trial (RCT). A pragmatic design will be been chosen to reflect how the Tute intervention is run in normal circumstances in schools. Consequently, teachers

will be given the freedom to select pupils for the study as they would do in normal teaching practice when deploying Tute as an intervention. The trial will involve 5 schools in cohort 1, with 120 Year 7 pupils block randomised based on pre-test performance into either the intervention or control group, controlling for any unknown hidden biases when conducting the final analyses. The design used blocked randomisation with pre-test Alfiesoft data ensuring the control and intervention groups are balanced for mathematics ability[31]. The trial will be designed as an aggregated trial, as it will not feasible to conduct a large scale RCT with sufficient power with the resources available in this study. The aggregated approach will allow the trial to be repeated with a minimum of two further cohorts required over the next academic year.

This is a waiting list design. This design is ethical as it does not deprive anyone of the treatment. It is less likely to demoralise schools / pupils in the controlled group than had there not been a waiting list. If the intervention is successful, the pupils will receive the intervention in the next academic year.

Full details of the background and trial design can be found within the protocol (Harrison, 2016).

Sample size

The focus of this trial will be on pupils who are performing at "below" or "expected" progress, therefore the sample size was based on this subgroup of children and limited to 12 pupil places in the TUTE virtual classroom. The initial cohort for the cumulative meta-analysis will involve 5 schools, with 24 pupils selected from Year 7. The total sample will be 120 pupils, 60 receiving the intervention and 60 continuing as business as usual. A recent EEF evaluation (Torgerson et al., 2014) found a positive effect size for small group tuition (weighted mean effect size[32] of 0.34). The aggregated study is powered to have an 80% chance of detecting an effect size of 0.34, with a sample of 360 pupils. However, through the prospective nature of a cumulative meta-analysis, repeating the trial and aggregating the results would allow for a more robust analysis of effectiveness.

The intra-class correlation rho = 0.15 accounts for the variance between schools, rho = 0.15 is selected as this is the recommended intra-class correlation for achievement in Mathematics (Tymms et al., 2011).

[31] Mathematics ability is defined by the topic assessed on Alfiesoft (fractions).
[32] Effect size used Cohen's d.

Randomisation

Randomisation will be conducted at pupil level in each school, carried out by an independent individual (independent teacher). Randomisation will be blinded to prevent selection bias as a recent systematic review of trials found studies using non-blinded assessors generated more optimistic effect estimates than blinded assessors (Hróbjartsson et al., 2014). The 24 pupils will be ranked in order of performance using Alfiesoft mathematics assessment pre-test (1 equating to best and 24 the least) and paired on ability. A coin will be used to select one of each successive pair of classes starting with 1&2 (then 3&4, etc.) forming the control and intervention group. Through blocking the performance ability as indicated by the Alfiesoft pre-assessment on mathematics performance this should result in an equal distribution across the groups.

Outcomes

Alfiesoft online assessments will be the main test used to determine mathematics outcomes. The online assessments will be self-marking, reducing the possibility of bias in the trial. The software is developed by AQA and used across many schools in the UK.

The primary outcome will be the attainment score on the fraction assessment created using the Alfiesoft system. The assessment consisted of 19 questions and 41 marks.

Data

Baseline data

Pupil baseline data for the intervention and control groups will be presented by trial arm to assess balance. Pupil level characteristics will be summarised by trail arm for both as randomised and included in the primary analysis. Information on the lessons from which the pupils were withdrawn is also summarised. No formal statistical testing to assess balance was conducted.

Outcome data

The online ALFIE assessments will be delivered on an agreed date prior to the start of the intervention and in the final week of the planned online lesson. The assessments were completed under 'exam conditions' with all pupils (control and intervention) participating at the same time.

Cleaning and formatting

Baseline data will be checked upon receipt for completeness and to ensure pupils are eligible to take part in the trial.

Analysis

Analysis will use intention to treat, therefore all pupils will be analysed in the group they were randomly allocated regardless of whether or not they received the intervention. Analyses will be conducted using R statistical software, using a package developed by Durham University which is commonly used in EEF evaluations. Effect sizes are presented relating to the analyses alongside 95% confidence intervals. The effect size is defined as:

$$g = \frac{\bar{x}_1 - \bar{x}_2}{s^*}$$

where the pooled standard deviation *s** is computed as:

$$s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

Trial Completion (CONSORT flow diagram)

A CONSORT diagram will be produced to show the flow of schools and pupils through the trial. This will include the number of pupils opting out of the trial.

Primary analysis

The primary objective of this study is will be to investigate the effectiveness of the Tute intervention on the mathematics skills of the pupils in the intervention group. The primary measure will be the Alfiesoft mathematics scores and will be analysed using a multi-level model. In the model, schools are random and the pre-test and post-test assigned as fixed effects. This method will be chosen as the design is using blocked randomisation, controlling for all unknown variables effecting mathematics scores. The standard 'effect' size was calculated as the difference between the scores for the intervention and control, divided by the pooled (average) post-test standard deviation of both groups.

The primary outcome measure is mathematics attainment and this will be marked blind by the Alfiesoft online assessment.

Secondary analyses

A secondary analysis using the data from the intervention group using a multi-level analysis will be completed to determine the impact of the intervention on pupil premium funded students. The gained scores from pre- post tests will be used and compared at school level, together with pooled standard deviation, to establish if the intervention had an impact on pupil premium students. Also, an analysis on the dosage and behaviour for learning will be include in the analysis.

Cumulative meta-analysis

Two additional cohorts in November 2016 and March 2017. The data will be combined with the feasibility study to model the development of a prospective cumulative meta-analysis using a number of models.

Compliance

Non-compliance will be summarised in terms of student attendance (number of lessons attended) with thresholds at 75% and 50%. In addition, the number of pupils attending 75% and 50% of sessions on time will be summarised. The impact of non-compliance (if this occurs) will be assessed using Complier Average Casual Effect (CACE), as outlined in the secondary analyses.

Appendix 7: Ethics Approval Letters

Durham University
School of Education

Shaped by the past, creating the future

25 September 2015

Wayne Harrison
PhD Education

w.d.harrison@durham.ac.uk

Dear Wayne

Online Maths Peer Tuition – Pilot study

I am pleased to inform you that your application for ethical approval for the above research has been approved by the School of Education Ethics Committee. May we take this opportunity to wish you good luck with your research.

P. M. Holmes

Dr. P. Holmes
Chair of School of Education Ethics Committee

Leazes Road
Durham, DH1 1TA
Telephone +44 (0)191 334 2000 Fax +44 (0)191 334 8311
www.dur.ac.uk/education

Durham University
School of Education

Shaped by the past, creating the future

19th January 2016

Wayne Harrison
PhD Education

w.d.harrison@durham.ac.uk

Dear Wayne

Online Maths Peer Tuition – Pilot study

Amended January 2016 as follows :

To repeat the online cross-age peer tuition project with schools over an 18 month data collection. The only changes are the dates of delivery and schools. The design, letters, opt-out forms, assessments are identical to the original application. The next trial will start 29th February 2016

I am pleased to inform you that your application for ethical approval for the above research has been approved by the School of Education Ethics Committee. May we take this opportunity to wish you good luck with your research.

P. M. Holmes

Dr. P. Holmes
Chair of School of Education Ethics Committee

Leazes Road
Durham, DH1 1TA
Telephone +44 (0)191 334 2000 Fax +44 (0)191 334 8311
www.durham.ac.uk/education

Durham University
School of Education

Shaped by the past, creating the future

29 April 2016

Wayne Harrison
PhD Education

w.d.harrison@durham.ac.uk

Dear Wayne

Online Maths Peer Tuition – Pilot study

Amended April 2016 as follows:

To repeat the online cross-age peer tuition project with schools over an 18 month data collection. The only changes are the dates of delivery and schools. The design, letters, opt-out forms, assessments are identical to the original application. The next trial will start 6th June 2016

I am pleased to inform you that your application for ethical approval for the above research has been approved by the School of Education Ethics Committee. May we take this opportunity to wish you good luck with your research.

P. M. Holmes

Dr. P. Holmes
Chair of School of Education Ethics Committee

Leazes Road
Durham, DH1 1TA
Telephone +44 (0)191 334 2000 Fax +44 (0)191 334 8311
www.durham.ac.uk/education

Durham University
School of Education

Shaped by the past, creating the future

4 May 2016

Wayne Harrison
w.d.harrison@durham.ac.uk

Dear Wayne

Tute online small group learning study

I am pleased to inform you that your application for ethical approval for the above research has been approved by the School of Education Ethics Committee. May we take this opportunity to wish you good luck with your research.

P. M. Holmes

Dr. P. Holmes
Chair of School of Education Ethics Committee

Leazes Road
Durham, DH1 1TA
Telephone +44 (0)191 334 8400 Fax +44 (0)191 334 8311
www.durham.ac.uk/education

Appendix 8: SPIRIT 2013 Checklist

The SPIRIT 2013 Checklist: Recommended items to address in a clinical trial protocol and related documents.

Table 1. **SPIRIT 2013 Checklist: Recommended Items to Address in a Clinical Trial Protocol and Related Documents***

Section/Item	Item Number	Description
Administrative information		
Title	1	Descriptive title identifying the study design, population, interventions, and, if applicable, trial acronym
Trial registration	2a	Trial identifier and registry name. If not yet registered, name of intended registry.
	2b	All items from the World Health Organization Trial Registration Data Set (**Appendix Table**, available at www.annals.org)
Protocol version	3	Date and version identifier
Funding	4	Sources and types of financial, material, and other support
Roles and responsibilities	5a	Names, affiliations, and roles of protocol contributors
	5b	Name and contact information for the trial sponsor
	5c	Role of study sponsor and funders, if any, in study design; collection, management, analysis, and interpretation of data; writing of the report; and the decision to submit the report for publication, including whether they will have ultimate authority over any of these activities
	5d	Composition, roles, and responsibilities of the coordinating center, steering committee, end point adjudication committee, data management team, and other individuals or groups overseeing the trial, if applicable (see item 21a for DMC)
Introduction		
Background and rationale	6a	Description of research question and justification for undertaking the trial, including summary of relevant studies (published and unpublished) examining benefits and harms for each intervention
	6b	Explanation for choice of comparators
Objectives	7	Specific objectives or hypotheses
Trial design	8	Description of trial design, including type of trial (e.g., parallel group, crossover, factorial, single group), allocation ratio, and framework (e.g., superiority, equivalence, noninferiority, exploratory)
Methods		
Participants, interventions, and outcomes		
Study setting	9	Description of study settings (e.g., community clinic, academic hospital) and list of countries where data will be collected. Reference to where list of study sites can be obtained
Eligibility criteria	10	Inclusion and exclusion criteria for participants. If applicable, eligibility criteria for study centers and individuals who will perform the interventions (e.g., surgeons, psychotherapists)
Interventions	11a	Interventions for each group with sufficient detail to allow replication, including how and when they will be administered
	11b	Criteria for discontinuing or modifying allocated interventions for a given trial participant (e.g., drug dose change in response to harms, participant request, or improving/worsening disease)
	11c	Strategies to improve adherence to intervention protocols, and any procedures for monitoring adherence (e.g., drug tablet return, laboratory tests)
	11d	Relevant concomitant care and interventions that are permitted or prohibited during the trial
Outcomes	12	Primary, secondary, and other outcomes, including the specific measurement variable (e.g., systolic blood pressure), analysis metric (e.g., change from baseline, final value, time to event), method of aggregation (e.g., median, proportion), and time point for each outcome. Explanation of the clinical relevance of chosen efficacy and harm outcomes is strongly recommended
Participant timeline	13	Time schedule of enrollment, interventions (including any run-ins and washouts), assessments, and visits for participants. A schematic diagram is highly recommended (**Figure**).
Sample size	14	Estimated number of participants needed to achieve study objectives and how it was determined, including clinical and statistical assumptions supporting any sample size calculations
Recruitment	15	Strategies for achieving adequate participant enrollment to reach target sample size
Assignment of interventions (for controlled trials)		
Allocation		
Sequence generation	16a	Method of generating the allocation sequence (e.g., computer-generated random numbers), and list of any factors for stratification. To reduce predictability of a random sequence, details of any planned restriction (e.g., blocking) should be provided in a separate document that is unavailable to those who enroll participants or assign interventions.
Allocation concealment mechanism	16b	Mechanism of implementing the allocation sequence (e.g., central telephone; sequentially numbered, opaque, sealed envelopes), describing any steps to conceal the sequence until interventions are assigned
Implementation	16c	Who will generate the allocation sequence, who will enroll participants, and who will assign participants to interventions
Blinding (masking)	17a	Who will be blinded after assignment to interventions (e.g., trial participants, care providers, outcome assessors, data analysts), and how
	17b	If blinded, circumstances under which unblinding is permissible, and procedure for revealing a participant's allocated intervention during the trial

Continued on following page

Table 1—Continued

Section/Item	Item Number	Description
Data collection, management, and analysis		
Data collection methods	18a	Plans for assessment and collection of outcome, baseline, and other trial data, including any related processes to promote data quality (e.g., duplicate measurements, training of assessors) and a description of study instruments (e.g., questionnaires, laboratory tests) along with their reliability and validity, if known. Reference to where data collection forms can be found, if not in the protocol.
	18b	Plans to promote participant retention and complete follow-up, including list of any outcome data to be collected for participants who discontinue or deviate from intervention protocols
Data management	19	Plans for data entry, coding, security, and storage, including any related processes to promote data quality (e.g., double data entry; range checks for data values). Reference to where details of data management procedures can be found, if not in the protocol.
Statistical methods	20a	Statistical methods for analyzing primary and secondary outcomes. Reference to where other details of the statistical analysis plan can be found, if not in the protocol.
	20b	Methods for any additional analyses (e.g., subgroup and adjusted analyses)
	20c	Definition of analysis population relating to protocol nonadherence (e.g., as-randomized analysis), and any statistical methods to handle missing data (e.g., multiple imputation)
Monitoring		
Data monitoring	21a	Composition of DMC; summary of its role and reporting structure; statement of whether it is independent from the sponsor and competing interests; and reference to where further details about its charter can be found, if not in the protocol. Alternatively, an explanation of why a DMC is not needed.
	21b	Description of any interim analyses and stopping guidelines, including who will have access to these interim results and make the final decision to terminate the trial
Harms	22	Plans for collecting, assessing, reporting, and managing solicited and spontaneously reported adverse events and other unintended effects of trial interventions or trial conduct
Auditing	23	Frequency and procedures for auditing trial conduct, if any, and whether the process will be independent from investigators and the sponsor
Ethics and dissemination		
Research ethics approval	24	Plans for seeking REC/IRB approval
Protocol amendments	25	Plans for communicating important protocol modifications (e.g., changes to eligibility criteria, outcomes, analyses) to relevant parties (e.g., investigators, RECs/IRBs, trial participants, trial registries, journals, regulators)
Consent or assent	26a	Who will obtain informed consent or assent from potential trial participants or authorized surrogates, and how (see item 32)
	26b	Additional consent provisions for collection and use of participant data and biological specimens in ancillary studies, if applicable
Confidentiality	27	How personal information about potential and enrolled participants will be collected, shared, and maintained in order to protect confidentiality before, during, and after the trial
Declaration of interests	28	Financial and other competing interests for principal investigators for the overall trial and each study site
Access to data	29	Statement of who will have access to the final trial data set, and disclosure of contractual agreements that limit such access for investigators
Ancillary and post-trial care	30	Provisions, if any, for ancillary and post-trial care, and for compensation to those who suffer harm from trial participation
Dissemination policy	31a	Plans for investigators and sponsor to communicate trial results to participants, health care professionals, the public, and other relevant groups (e.g., via publication, reporting in results databases, or other data-sharing arrangements), including any publication restrictions
	31b	Authorship eligibility guidelines and any intended use of professional writers
	31c	Plans, if any, for granting public access to the full protocol, participant-level data set, and statistical code
Appendices		
Informed consent materials	32	Model consent form and other related documentation given to participants and authorized surrogates
Biological specimens	33	Plans for collection, laboratory evaluation, and storage of biological specimens for genetic or molecular analysis in the current trial and for future use in ancillary studies, if applicable

DMC = data monitoring committee; IRB = institutional review board; REC = research ethics committee; SPIRIT = Standard Protocol Items: Recommendations for Interventional Trials.

* It is strongly recommended that this checklist be read in conjunction with the SPIRIT 2013 Explanation and Elaboration (31) for important clarification on the items. Amendments to the protocol should be tracked and dated. The SPIRIT checklist is copyrighted by the SPIRIT Group and is reproduced with permission.

(Chan et al., 2013)

Appendix 9 – R Code for Study One

Cohort 1

```
> setwd("C:/Users/wdhar/OneDrive/Desktop/PhD drafts/Chapter 5 Trial 1 Onli
ne peer tuition/Data WH")
> wdat <- read.csv("wdata1.csv", header=T); wdat
  ï..Pupil.Code Gender School Group.Size Role PP pret post
1         1      M      0      01:04   PT 0   15   17
2         2      F      0      01:01   PT 0   7   17
3         3      F      0      01:02   PT 0   11   17
4         4      F      0      01:01   PT 0   17   17
5         5      M      0      01:02   PT 0   10   15
6         6      M      0      01:01   PT 0   13   9
7         7      M      0      01:01   PT 1   9   14
8         9      M      0      01:02   PT 1   13   15
9         10      M      0      01:01   PT 0   14   15
10        11      M      0      01:04   PT 0   17   16
11        12      F      0      01:01   PT 1   8   15
12        13      M      0      01:01   PT 0   17   16
13        14      M      0      01:01   PT 0   14   17
14        15      M      0      01:01   PT 0   13   15
15        16      M      0      01:01   PT 0   15   17
16        17      M      0      01:01   PT 1   5   15
17        18      F      0      01:02   PT 0   10   16
18        19      F      0      01:04   PT 0   15   17
19        20      M      0      01:04   PT 0   11   16
20        21      M      0      01:02   PT 0   8   15
21        22      M      0      01:01   PT 1   11   11
22        23      F      1      01:04   P 0   3   6
23        24      M      1      01:01   P 0   9   17
24        25      M      1      01:02   P 1   3   10
25        26      M      1      01:01   P 0   3   15
26        27      M      1      01:04   P 0   2   14
27        28      F      1      01:01   P 0   2   12
28        29      F      1      01:04   P 0   4   17
29        30      F      1      01:02   P 0   4   12
30        31      F      1      01:02   P 1   2   12
31        32      F      1      01:02   P 0   5   15
32        33      F      1      01:01   P 0   4   11
33        34      M      2      01:04   P 0   2   4
34        35      M      2      01:02   P 0   7   16
35        36      M      2      01:01   P 1   2   7
36        37      M      2      01:02   P 0   4   12
37        38      F      2      01:04   P 0   3   7
38        39      M      2      01:02   P 1   1   3
39        40      F      2      01:04   P 1   7   10
40        41      M      2      01:04   P 0   2   7
41        42      M      2      01:01   P 0   2   9
42        44      F      2      01:01   P 0   4   11
43        45      M      2      01:04   P 0   3   10
44        46      M      2      01:01   P 1   10   13
45        47      F      2      01:04   P 0   5   7
46        48      F      2      01:01   P 0   7   11
47        49      F      2      01:04   P 0   1   7
48        50      F      2      01:02   P 0   2   7
49        51      F      2      01:01   P 0   1   14
50        52      F      2      01:01   P 1   2   8
51        53      M      2      01:02   P 0   2   10
52        54      F      2      01:04   P 0   2   9
53        55      M      2      01:01   P 0   3   8
54        56      F      2      01:02   P 0   3   6
55        57      F      2      01:01   P 1   0   9
56        58      M      2      01:02   P 0   8   13
> wdat$female <- NA; wdat$female[wdat$Gender %in% "M"] <- 0; wdat$female[w
dat$Gender %in% "F"] <- 1; wdat$female
```

```
 [1] 0 1 1 1 0 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 1 1 1 1 1 1 0 0 0
0 1 0 1 0 0 1 0 0 1 1 1 1 1 1 0 1 0 1 1 0
> wdat$t <- NA; wdat$t[wdat$Group.Size %in% "01:04"] <- 2; wdat$t[wdat$Gro
up.Size %in% "01:02"] <- 1; wdat$t[wdat$Group.Size %in% "01:01"] <- 0; wda
t$t
 [1] 2 0 1 0 1 0 0 1 0 2 0 0 0 0 0 0 1 2 2 1 0 2 0 1 0 2 0 2 1 1 1 0 2 1 0
1 2 1 2 2 0 0 2 0 2 0 2 1 0 0 1 2 0 1 0 1
> wdat$post <- wdat$post; wdat$pret <- wdat$pret
> head(wdat)
 ï..Pupil.Code Gender School Group.Size Role PP pret post female t
1         1      M      0     01:04   PT 0   15   17     0 2
2         2      F      0     01:01   PT 0    7   17     1 0
3         3      F      0     01:02   PT 0   11   17     1 1
4         4      F      0     01:01   PT 0   17   17     1 0
5         5      M      0     01:02   PT 0   10   15     0 1
6         6      M      0     01:01   PT 0   13    9     0 0
> wdat <- na.omit(wdat[c("post", "t", "pret", "School", "female")]); dim(w
dat)
[1] 56 5
> head(wdat)
 post t pret School female
1   17 2   15      0      0
2   17 0    7      0      1
3   17 1   11      0      1
4   17 0   17      0      1
5   15 1   10      0      0
6    9 0   13      0      0
> wd.1to4 <- subset(wdat, wdat$t==2 | wdat$t==0); dim(wd.1to4)
[1] 40 5
> wd.1to2 <- subset(wdat, wdat$t==1 | wdat$t==0); dim(wd.1to2)
[1] 41 5
> wd.2.4 <- subset(wdat, wdat$t==2 | wdat$t==1); dim(wd.2.4)
[1] 31 5
> wd.1to2$t
 [1] 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 1 0 0 1 0 0 1 1 1 0 1 0 1 1 0 0 0 0 1 0
0 1 0 1 0 1
> wd.1to2$post
 [1] 17 17 17 15 9 14 15 15 15 16 17 15 17 15 16 15 11 17 10 15 12 12 12 1
5 11 16 7 12 3 9 11 13 11 7 14 8 10 8 6
[40] 9 13
> efs2(wd.1to4$post, wd.1to4$t)
   cd cd.lb cd.ub  g g.lb g.ub  d d.lb d.ub
1 -0.51 -1.16 0.14 -0.5 -1.14 0.14 -0.6 -1.25 0.05
> efs(wd.1to2$post, wd.1to2$t)
   cd cd.lb cd.ub   g g.lb g.ub   d d.lb d.ub
1 -0.22 -0.85 0.41 -0.22 -0.83 0.4 -0.24 -0.87 0.39
> ancovaES2(wd.1to4$post, wd.1to4$t, wd.1to4$pret)
  cd cd.lb cd.ub  g g.lb g.ub beta sd.pool
1 0.16 -0.48  0.8 0.16 -0.47 0.78 0.47  2.96
> ancovaES(wd.1to2$post, wd.1to2$t, wd.1to2$pret)
  cd cd.lb cd.ub  g g.lb g.ub beta sd.pool
1 0.04 -0.59 0.67 0.04 -0.58 0.65 0.11  2.78
> install.packages("nlme")
Error in install.packages : Updating loaded packages
> install.packages("nlme")
Installing package into 'C:/Users/wdhar/OneDrive/Documents/R/win-library/3
.6'
(as 'lib' is unspecified)
Warning in install.packages :
 package 'nlme' is in use and will not be installed
> library(nlme)
> lme <- lme(post ~ t + pret, random = ~ 1|School, data=wd.1to4, na.action
=na.exclude); summary(lme)
Linear mixed-effects model fit by REML
 Data: wd.1to4
    AIC   BIC  logLik
 202.6157 210.6703 -96.30783

Random effects:
```

```
 Formula: ~1 | School
    (Intercept) Residual
StdDev:  2.107065 2.571326

Fixed effects: post ~ t + pret
               Value Std.Error DF  t-value p-value
(Intercept) 10.635526 1.5657058 35 6.792800 0.0000
t       -0.505833 0.4291429 35 -1.178706 0.2465
pret     0.322204 0.1251090 35 2.575390 0.0144
 Correlation:
   (Intr) t
t  -0.190
pret -0.525 -0.036

Standardized Within-Group Residuals:
    Min      Q1     Med      Q3     Max
-2.55668983 -0.43598664 0.07932269 0.62925885 2.01934352

Number of Observations: 40
Number of Groups: 3
> beta.1to4 <- data.frame(intervals(lme, which="fixed")$fixed)[2,2]; beta.
1to4
[1] -0.5058332
> var.b.1to4 <- as.numeric(getVarCov(lme)); var.b.1to4
[1] 4.439724
> var.w.1to4 <- lme$sigma^2; var.w.1to4
[1] 6.611718
> var.tt.1to4 <- var.w.1to4 + var.b.1to4; var.tt.1to4
[1] 11.05144
> icc.1to4 <- round(as.numeric(getVarCov(lme))/(as.numeric(getVarCov(lme))
+lme$sigma^2),2); icc.1to4
[1] 0.4
> length(unique(wd.1to4$School))
[1] 3
> tt.1to4 <- g.total2(var.tt.1to4, beta.1to4, icc.1to4, wd.1to4$t, wd.1to4
$School); tt.1to4
   g  LB  UB
1 -0.14 -1.34 1.06
> wth.1to4 <- g.within2(var.w.1to4, beta.1to4, icc.1to4, wd.1to4$t, wd.1to
4$School); wth.1to4
   g  LB  UB
1 -0.2 -1.75 1.36
> btw.1to4 <- g.between2(var.b.1to4, beta.1to4, icc.1to4, wd.1to4$t, wd.1t
o4$School); btw.1to4
   g  LB  UB
1 -0.27 -5.73 5.18
> lme <- lme(post ~ t + pret, random = ~ 1|School, data=wd.1to2, na.action
=na.exclude); summary(lme)
Linear mixed-effects model fit by REML
 Data: wd.1to2
    AIC   BIC  logLik
 204.7679 212.9558 -97.38395

Random effects:
 Formula: ~1 | School
    (Intercept) Residual
StdDev:  1.486005 2.563706

Fixed effects: post ~ t + pret
               Value Std.Error DF t-value p-value
(Intercept) 10.051535 1.3119444 36 7.661556 0.0000
t       0.028950 0.8370108 36 0.034587 0.9726
pret     0.392807 0.1221705 36 3.215238 0.0028
 Correlation:
   (Intr) t
t  -0.332
pret -0.636 0.121

Standardized Within-Group Residuals:
```

```
    Min      Q1     Med      Q3     Max
-2.56566477 -0.44811717 0.06616338 0.54078955 1.93800063

Number of Observations: 41
Number of Groups: 3
> beta.1to2 <- data.frame(intervals(lme, which="fixed")$fixed)[2,2]; beta.
1to2
[1] 0.02894963
> var.b.1to2 <- as.numeric(getVarCov(lme)); var.b.1to2
[1] 2.208212
> var.w.1to2 <- lme$sigma^2; var.w.1to2
[1] 6.572591
> var.tt.1to2 <- var.w.1to2 + var.b.1to2; var.tt.1to2
[1] 8.780803
> icc.1to2 <- round(as.numeric(getVarCov(lme))/(as.numeric(getVarCov(lme))
+lme$sigma^2),2); icc.1to2
[1] 0.25
> length(unique(wd.1to2$School))
[1] 3
> tt.1to2 <- g.total(var.tt.1to2, beta.1to2, icc.1to2, wd.1to2$t, wd.1to2$
School); tt.1to2
   g  LB  UB
1 0.01 -0.99 1.01
> wth.1to2 <- g.within(var.w.1to2, beta.1to2, icc.1to2, wd.1to2$t, wd.1to2
$School); wth.1to2
   g  LB  UB
1 0.01 -1.15 1.17
> btw.1to2 <- g.between1(var.b.1to2, beta.1to2, icc.1to2, wd.1to2$t, wd.1t
o2$School); btw.1to2
   g  LB  UB
1 0.02 -5.89 5.94
```

Cohort 2

```
> setwd("C:/Users/wdhar/OneDrive/Desktop/PhD drafts/Chapter 5 Trial 1 Onli
ne peer tuition/Data WH")
> wdat <- read.csv("wdata2.csv", header=T); wdat
  ï..Pupil.Code Gender School Group.Size Role PP pret post
1         81      M      5     01:01   PT 1    5   12
2         82      M      5     01:01   PT 0    5   12
3         82      F      5     01:01   PT 1    8   12
4         83      F      5     01:01   PT 0    4   7
5         84      F      5     01:04   PT 0    7   11
6         85      F      5     01:02   PT 0    5   10
7         86      M      5     01:02   PT 0    4   14
8         87      M      6     01:04   P 0     2   7
9         88      M      6     01:01   P 1     3   8
10        89      M      6     01:04   P 0     1   5
11        90      F      6     01:02   P 1     2   7
12        91      F      6     01:01   P 1     2   9
13        92      M      6     01:02   P 1     1   8
14        93      M      6     01:04   P 0     2   9
15        94      F      6     01:02   P 1     2   8
16        95      F      6     01:01   P 1     0   6
17        96      M      6     01:01   P 0     2   7
18        97      F      6     01:02   P 0     2   9
19        99      M      5     01:01   PT 0    6   13
20       100      M      5     01:01   PT 0    8   12
21       101      F      5     01:02   PT 1    8   12
22       102      F      5     01:01   PT 0    8   11
23       103      F      5     01:04   PT 0    7   10
24       104      F      5     01:01   PT 1    6   11
25       105      M      5     01:02   PT 0    9   12
26       106      M      6     01:04   P 0     2   6
27       107      M      6     01:01   P 0     2   5
28       108      M      6     01:04   P 1     1   6
29       109      M      6     01:02   P 1     2   7
30       110      F      6     01:01   P 0     1   6
31       111      F      6     01:02   P 1     1   7
32       112      M      6     01:04   P 0     2   5
33       113      F      6     01:02   P 0     1   4
34       114      M      6     01:01   P 0     1   6
35       115      F      6     01:01   P 0     2   5
36       116      F      6     01:02   P 1     2   8
37       117      M      6     01:04   P 0     1   6
> wdat$female <- NA; wdat$female[wdat$Gender %in% "M"] <- 0; wdat$female[w
dat$Gender %in% "F"] <- 1; wdat$female
 [1] 0 0 1 1 1 1 0 0 0 0 1 1 0 0 1 1 0 1 0 0 1 1 1 1 0 0 0 0 0 1 1 0 1 0 1
1 0
> wdat$t <- NA; wdat$t[wdat$Group.Size %in% "01:04"] <- 2; wdat$t[wdat$Gro
up.Size %in% "01:02"] <- 1; wdat$t[wdat$Group.Size %in% "01:01"] <- 0; wda
t$t
 [1] 0 0 0 0 2 1 1 2 0 2 1 0 1 2 1 0 0 1 0 0 1 0 2 0 1 2 0 2 1 0 1 2 1 0 0
1 2
> wdat$post <- wdat$post; wdat$pret <- wdat$pret
> head(wdat)
 ï..Pupil.Code Gender School Group.Size Role PP pret post female t
1         81      M      5     01:01   PT 1    5   12      0 0
2         82      M      5     01:01   PT 0    5   12      0 0
3         82      F      5     01:01   PT 1    8   12      1 0
4         83      F      5     01:01   PT 0    4   7     1 0
5         84      F      5     01:04   PT 0    7   11      1 2
6         85      F      5     01:02   PT 0    5   10      1 1
> wdat <- na.omit(wdat[c("post", "t", "pret", "School", "female")]); dim(w
dat)
[1] 37 5
> head(wdat)
 post t pret School female
1   12 0    5      5      0
2   12 0    5      5      0
3   12 0    8      5      1
```

```
4  7 0  4   5   1
5  11 2  7   5   1
6  10 1  5   5   1
> wd.1to4 <- subset(wdat, wdat$t==2 | wdat$t==0); dim(wd.1to4)
[1] 25 5
> wd.1to2 <- subset(wdat, wdat$t==1 | wdat$t==0); dim(wd.1to2)
[1] 28 5
> wd.2.4 <- subset(wdat, wdat$t==2 | wdat$t==1); dim(wd.2.4)
[1] 21 5
> wd.1to2$t
 [1] 0 0 0 0 1 1 0 1 0 1 1 0 0 1 0 0 1 0 0 1 0 1 0 1 1 0 0 1
> wd.1to2$post
 [1] 12 12 12 7 10 14 8 7 9 8 8 6 7 9 13 12 12 11 11 12 5 7 6 7 4 6 5 8
> efs2(wd.1to4$post, wd.1to4$t)
   cd cd.lb cd.ub   g g.lb g.ub   d d.lb d.ub
1 -0.61 -1.45 0.22 -0.59 -1.4 0.21 -0.57 -1.4 0.27
> efs(wd.1to2$post, wd.1to2$t)
   cd cd.lb cd.ub   g g.lb g.ub   d d.lb d.ub
1 -0.01 -0.76 0.73 -0.01 -0.74 0.71 -0.01 -0.76 0.73
> ancovaES2(wd.1to4$post, wd.1to4$t, wd.1to4$pret)
  cd cd.lb cd.ub  g g.lb g.ub beta sd.pool
1 0.62 -0.22 1.45 0.6 -0.21 1.4 0.88  1.43
> ancovaES(wd.1to2$post, wd.1to2$t, wd.1to2$pret)
  cd cd.lb cd.ub  g g.lb g.ub beta sd.pool
1 0.32 -0.43 1.08 0.31 -0.42 1.04 0.55  1.69
> install.packages("nlme")
Error in install.packages : Updating loaded packages
> install.packages("nlme")
Installing package into 'C:/Users/wdhar/OneDrive/Documents/R/win-library/3
.6'
(as 'lib' is unspecified)
Warning in install.packages :
 package 'nlme' is in use and will not be installed
> library(nlme)
> lme <- lme(post ~ t + pret, random = ~ 1|School, data=wd.1to4, na.action
=na.exclude); summary(lme)
Linear mixed-effects model fit by REML
 Data: wd.1to4
    AIC   BIC  logLik
 99.46573 104.9209 -44.73286

Random effects:
 Formula: ~1 | School
    (Intercept) Residual
StdDev:  0.8476523 1.409906

Fixed effects: post ~ t + pret
        Value Std.Error DF  t-value p-value
(Intercept) 5.907776 1.0381351 21 5.690759 0.0000
t      -0.284317 0.3030954 21 -0.938046 0.3589
pret     0.753581 0.1957744 21 3.849230 0.0009
 Correlation:
   (Intr) t
t  -0.217
pret -0.743 0.026

Standardized Within-Group Residuals:
    Min     Q1     Med     Q3     Max
-1.63163827 -0.50667485 -0.02594443 0.33376831 1.79590323

Number of Observations: 25
Number of Groups: 2
> beta.1to4 <- data.frame(intervals(lme, which="fixed")$fixed)[2,2]; beta.
1to4
[1] -0.2843173
> var.b.1to4 <- as.numeric(getVarCov(lme)); var.b.1to4
[1] 0.7185144
> var.w.1to4 <- lme$sigma^2; var.w.1to4
[1] 1.987836
```

```
> var.tt.1to4 <- var.w.1to4 + var.b.1to4; var.tt.1to4
[1] 2.70635
> icc.1to4 <- round(as.numeric(getVarCov(lme))/(as.numeric(getVarCov(lme))
+lme$sigma^2),2); icc.1to4
[1] 0.27
> length(unique(wd.1to4$School))
[1] 2
> tt.1to4 <- g.total2(var.tt.1to4, beta.1to4, icc.1to4, wd.1to4$t, wd.1to4
$School); tt.1to4
   g  LB  UB
1 -0.16 -1.46 1.14
> wth.1to4 <- g.within2(var.w.1to4, beta.1to4, icc.1to4, wd.1to4$t, wd.1to
4$School); wth.1to4
   g  LB  UB
1 -0.2 -1.72 1.32
> btw.1to4 <- g.between2(var.b.1to4, beta.1to4, icc.1to4, wd.1to4$t, wd.1t
o4$School); btw.1to4
   g  LB  UB
1 -0.42 -5.42 4.57
> lme <- lme(post ~ t + pret, random = ~ 1|School, data=wd.1to2, na.action
=na.exclude); summary(lme)
Linear mixed-effects model fit by REML
 Data: wd.1to2
    AIC   BIC  logLik
 114.7474 120.8418 -52.37369

Random effects:
 Formula: ~1 | School
    (Intercept) Residual
StdDev:  2.091301 1.534814

Fixed effects: post ~ t + pret
        Value Std.Error DF t-value p-value
(Intercept) 7.274804 1.7586441 24 4.136598 0.0004
t      0.691048 0.5943086 24 1.162776 0.2564
pret    0.406399 0.2212342 24 1.836963 0.0786
 Correlation:
   (Intr) t
t  -0.128
pret -0.495 -0.026

Standardized Within-Group Residuals:
    Min     Q1     Med     Q3     Max
-2.12426147 -0.42625881 0.02372831 0.58403332 1.98630397

Number of Observations: 28
Number of Groups: 2
> beta.1to2 <- data.frame(intervals(lme, which="fixed")$fixed)[2,2]; beta.
1to2
[1] 0.6910479
> var.b.1to2 <- as.numeric(getVarCov(lme)); var.b.1to2
[1] 4.373538
> var.w.1to2 <- lme$sigma^2; var.w.1to2
[1] 2.355653
> var.tt.1to2 <- var.w.1to2 + var.b.1to2; var.tt.1to2
[1] 6.729191
> icc.1to2 <- round(as.numeric(getVarCov(lme))/(as.numeric(getVarCov(lme))
+lme$sigma^2),2); icc.1to2
[1] 0.65
> length(unique(wd.1to2$School))
[1] 2
> tt.1to2 <- g.total(var.tt.1to2, beta.1to2, icc.1to2, wd.1to2$t, wd.1to2$
School); tt.1to2
   g  LB  UB
1 0.22 -1.47 1.91
> wth.1to2 <- g.within(var.w.1to2, beta.1to2, icc.1to2, wd.1to2$t, wd.1to2
$School); wth.1to2
   g  LB UB
1 0.45 -2.4 3.3
```

```
> btw.1to2 <- g.between1(var.b.1to2, beta.1to2, icc.1to2, wd.1to2$t, wd.1t
o2$School); btw.1to2
   g  LB  UB
1 0.34 -3.75 4.44
```

Cohort 3

Analysis cohort 3

```
> setwd("C:/Users/wdhar/OneDrive/Desktop/PhD drafts/Chapter 5 Trial 1 Onli
ne peer tuition/Data WH")
> wdat <- read.csv("wdata3.csv", header=T); wdat
  ï..Pupil.Code Gender School Group.Size Role PP pret post
1       118    F    7   01:01  P 0  10  12
2       119    M    7   01:01  P 0  2  3
3       120    M    7   01:04  P 1  3  6
4       121    M    7   01:04  P 0  1  5
5       122    M    7   01:02  P 0  2  5
6       123    M    7   01:02  P 0  14  14
7       124    F    7   01:04  P 0  6  9
8       125    M    7   01:01  P 1  6  8
9       126    M    7   01:02  P 1  3  8
10       127    F    7   01:01  P 0  2  3
11       128    M    7   01:02  P 0  2  4
12       129    M    7   01:04  P 0  2  5
13       130    F    8   01:02  P 1  6  8
14       131    M    8   01:04  P 0  10  13
15       132    F    8   01:01  P 0  6  9
16       133    F    8   01:02  P 0  6  10
17       134    M    8   01:01  P 0  5  7
18       135    F    8   01:01  P 0  5  8
19       136    M    8   01:04  P 1  4  8
20       137    F    8   01:02  P 0  5  6
21       138    F    8   01:04  P 0  9  12
22       139    M    8   01:02  P 0  11  13
23       140    M    8   01:04  P 1  6  6
24       141    M    8   01:01  P 0  12  14
25       142    M    0   01:01  PT 0  12  15
26       143    F    0   01:01  PT 0  14  15
27       144    F    0   01:02  PT 0  13  14
28       145    M    0   01:01  PT 0  14  16
29       146    M    0   01:02  PT 1  12  16
30       147    F    0   01:04  PT 0  10  13
31       148    M    0   01:01  PT 0  14  14
32       149    F    0   01:02  PT 0  15  15
33       150    M    0   01:01  PT 1  11  13
34       151    M    0   01:04  PT 0  12  14
35       152    F    0   01:02  PT 0  12  15
36       153    F    0   01:01  PT 1  14  15
37       154    F    0   01:01  PT 0  12  14
38       155    M    0   01:01  PT 0  11  13
39       156    F    9   01:04  PT 0  12  14
40       157    F    9   01:01  PT 0  14  15
41       158    M    9   01:02  PT 0  15  15
42       159    M    9   01:01  PT 0  16  16
43       160    F    9   01:02  PT 0  15  15
44       161    M    9   01:01  PT 0  11  14
45       162    M    9   01:01  PT 0  8  13
46       163    M    9   01:01  PT 1  13  15
47       164    M    10   01:01  P 0  1  4
48       165    F    10   01:02  P 0  1  5
49       166    F    10   01:01  P 0  6  7
50       167    F    10   01:01  P 0  4  7
51       168    F    10   01:02  P 0  8  11
52       169    F    10   01:01  P 1  2  6
53       170    F    10   01:04  P 0  5  8
54       171    F    10   01:02  P 0  2  4
55       173    M    10   01:04  P 0  2  2
56       174    F    10   01:04  P 0  2  7
57       175    M    10   01:04  P 0  5  8
58       176    F    10   01:02  P 0  3  5
> wdat$female <- NA; wdat$female[wdat$Gender %in% "M"] <- 0; wdat$female[w
dat$Gender %in% "F"] <- 1; wdat$female
```

```
 [1] 1 0 0 0 0 0 1 0 0 1 0 0 1 0 1 1 0 1 0 1 1 0 0 0 0 1 1 0 0 1 0 1 0 0 1
1 1 0 1 1 0 0 1 0 0 0 0 0 1 1 1 1 1 1 1 0 1 0 1
> wdat$t <- NA; wdat$t[wdat$Group.Size %in% "01:04"] <- 2; wdat$t[wdat$Gro
up.Size %in% "01:02"] <- 1; wdat$t[wdat$Group.Size %in% "01:01"] <- 0; wda
t$t
 [1] 0 0 2 2 1 1 2 0 1 0 1 2 1 2 0 1 0 0 2 1 2 1 2 0 0 0 1 0 1 2 0 1 0 2 1
0 0 0 2 0 1 0 1 0 0 0 0 1 0 0 1 0 2 1 2 2 2 1
> wdat$post <- wdat$post; wdat$pret <- wdat$pret
> head(wdat)
 ï..Pupil.Code Gender School Group.Size Role PP pret post female t
1       118      F      7     01:01    P 0   10   12      1 0
2       119      M      7     01:01    P 0    2    3      0 0
3       120      M      7     01:04    P 1    3    6      0 2
4       121      M      7     01:04    P 0    1    5      0 2
5       122      M      7     01:02    P 0    2    5      0 1
6       123      M      7     01:02    P 0   14   14      0 1
> wdat <- na.omit(wdat[c("post", "t", "pret", "School", "female")]); dim(w
dat)
[1] 58 5
> head(wdat)
 post t pret School female
1   12 0   10      7      1
2    3 0    2      7      0
3    6 2    3      7      0
4    5 2    1      7      0
5    5 1    2      7      0
6   14 1   14      7      0
> wd.1to4 <- subset(wdat, wdat$t==2 | wdat$t==0); dim(wd.1to4)
[1] 40 5
> wd.1to2 <- subset(wdat, wdat$t==1 | wdat$t==0); dim(wd.1to2)
[1] 43 5
> wd.2.4 <- subset(wdat, wdat$t==2 | wdat$t==1); dim(wd.2.4)
[1] 33 5
> wd.1to2$t
 [1] 0 0 1 1 0 1 0 1 1 0 1 0 0 1 1 0 0 0 1 0 1 0 1 0 1 0 0 0 0 1 0 1 0 0 0
0 1 0 0 1 0 1 1
> wd.1to2$post
 [1] 12 3 5 14 8 8 3 4 8 9 10 7 8 6 13 14 15 15 14 16 16 14 15 13 15 15 14
13 15 15 16 15 14 13 15 4 5 7 7
[40] 11 6 4 5
> efs2(wd.1to4$post, wd.1to4$t)
   cd cd.lb cd.ub   g g.lb g.ub   d d.lb d.ub
1 -0.58 -1.23 0.08 -0.56 -1.2 0.08 -0.55 -1.2 0.1
> efs(wd.1to2$post, wd.1to2$t)
  cd cd.lb cd.ub   g g.lb g.ub  d d.lb d.ub
1 -0.2 -0.81 0.41 -0.19 -0.79 0.4 -0.2 -0.81 0.41
> ancovaES2(wd.1to4$post, wd.1to4$t, wd.1to4$pret)
  cd cd.lb cd.ub  g g.lb g.ub beta sd.pool
1 0.77 0.11 1.43 0.75 0.1 1.4 0.9  1.18
> ancovaES(wd.1to2$post, wd.1to2$t, wd.1to2$pret)
   cd cd.lb cd.ub   g g.lb g.ub beta sd.pool
1 -0.05 -0.66 0.56 -0.05 -0.64 0.55 -0.06  1.18
> install.packages("nlme")
Installing package into 'C:/Users/wdhar/OneDrive/Documents/R/win-library/3
.6'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/nlme_3.1-141.
zip'
Content type 'application/zip' length 2375127 bytes (2.3 MB)
downloaded 2.3 MB

package 'nlme' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\wdhar\AppData\Local\Temp\RtmpUltXFI\downloaded_packages
> library(nlme)
> lme <- lme(post ~ t + pret, random = ~ 1|School, data=wd.1to4, na.action
=na.exclude); summary(lme)
Linear mixed-effects model fit by REML
```

```
 Data: wd.1to4
    AIC    BIC  logLik
 140.9892 149.0438 -65.49459

Random effects:
 Formula: ~1 | School
     (Intercept) Residual
StdDev: 3.642752e-05 1.177413

Fixed effects: post ~ t + pret
        Value Std.Error DF  t-value p-value
(Intercept) 2.8975752 0.4581742 33 6.324178 0.0000
t      0.2005613 0.2035970 33 0.985089 0.3317
pret    0.9047139 0.0436698 33 20.717130 0.0000
 Correlation:
   (Intr) t
t  -0.579
pret -0.858 0.329

Standardized Within-Group Residuals:
    Min     Q1    Med     Q3    Max
-2.6397913 -0.3879273 0.1509520 0.5076233 2.4330571

Number of Observations: 40
Number of Groups: 5
> beta.1to4 <- data.frame(intervals(lme, which="fixed")$fixed)[2,2]; beta.
1to4
[1] 0.2005613
> var.b.1to4 <- as.numeric(getVarCov(lme)); var.b.1to4
[1] 1.326965e-09
> var.w.1to4 <- lme$sigma^2; var.w.1to4
[1] 1.386302
> var.tt.1to4 <- var.w.1to4 + var.b.1to4; var.tt.1to4
[1] 1.386302
> icc.1to4 <- round(as.numeric(getVarCov(lme))/(as.numeric(getVarCov(lme))
+lme$sigma^2),2); icc.1to4
[1] 0
> length(unique(wd.1to4$School))
[1] 5
> tt.1to4 <- g.total2(var.tt.1to4, beta.1to4, icc.1to4, wd.1to4$t, wd.1to4
$School); tt.1to4
   g  LB  UB
1 0.17 -0.47 0.81
> wth.1to4 <- g.within2(var.w.1to4, beta.1to4, icc.1to4, wd.1to4$t, wd.1to
4$School); wth.1to4
   g  LB  UB
1 0.17 -0.47 0.81
> btw.1to4 <- g.between2(var.b.1to4, beta.1to4, icc.1to4, wd.1to4$t, wd.1t
o4$School); btw.1to4
  g LB UB
1 Inf NaN Inf
> lme <- lme(post ~ t + pret, random = ~ 1|School, data=wd.1to2, na.action
=na.exclude); summary(lme)
Linear mixed-effects model fit by REML
 Data: wd.1to2
    AIC    BIC  logLik
 149.9316 158.376 -69.96582

Random effects:
 Formula: ~1 | School
     (Intercept) Residual
StdDev: 5.500777e-05 1.182354

Fixed effects: post ~ t + pret
        Value Std.Error DF t-value p-value
(Intercept) 3.275304 0.4114280 36 7.96082 0.0000
t      -0.058520 0.3671937 36 -0.15937 0.8743
pret     0.862744 0.0374091 36 23.06243 0.0000
 Correlation:
```

```
   (Intr) t
t  -0.449
pret -0.818 0.096

Standardized Within-Group Residuals:
    Min      Q1     Med      Q3     Max
-1.69221032 -0.79694597 0.04882429 0.50514135 2.38739293

Number of Observations: 43
Number of Groups: 5
> beta.1to2 <- data.frame(intervals(lme, which="fixed")$fixed)[2,2]; beta.
1to2
[1] -0.05851955
> var.b.1to2 <- as.numeric(getVarCov(lme)); var.b.1to2
[1] 3.025855e-09
> var.w.1to2 <- lme$sigma^2; var.w.1to2
[1] 1.397961
> var.tt.1to2 <- var.w.1to2 + var.b.1to2; var.tt.1to2
[1] 1.397961
> icc.1to2 <- round(as.numeric(getVarCov(lme))/(as.numeric(getVarCov(lme))
+lme$sigma^2),2); icc.1to2
[1] 0
> length(unique(wd.1to2$School))
[1] 5
> tt.1to2 <- g.total(var.tt.1to2, beta.1to2, icc.1to2, wd.1to2$t, wd.1to2$
School); tt.1to2
   g  LB  UB
1 -0.05 -0.66 0.56
> wth.1to2 <- g.within(var.w.1to2, beta.1to2, icc.1to2, wd.1to2$t, wd.1to2
$School); wth.1to2
   g  LB  UB
1 -0.05 -0.66 0.56
> btw.1to2 <- g.between1(var.b.1to2, beta.1to2, icc.1to2, wd.1to2$t, wd.1t
o2$School); btw.1to2
   g  LB UB
1 -Inf -Inf NaN
```

Appendix 10: R Code for Study Two

R code for the analysis

```
> dat <- read.csv("wh.csv"); dat

 school pre post intervention dose gender

1   0  3  4     0  0   0

2   0  7  NA    0  0   0

3   0 12  13    0  0   1

4   0  4  9     0  0   1

5   0  5  8     0  0   0

6   0  2  5     0  0   1

7   0  8  3     0  0   1

8   0  6  5     0  0   0

9   0  9  8     0  0   0

10  0  6  NA    0  0   1

11  0  3  3     0  0   1

12  0  5  7     0  0   1

13  1  4  8     0  0   1

14  1  7  9     0  0   1

15  1  7  7     0  0   1

16  1  9  9     0  0   1

17  1 10  8     0  0   0

18  1 10  12    0  0   1

19  1 11  9     0  0   1

20  1 12  11    0  0   1

21  1 12  9     0  0   1
```

22 1 15 11 0 0 0

23 1 16 10 0 0 1

24 1 17 13 0 0 1

25 2 11 12 0 0 0

26 2 15 13 0 0 0

27 2 16 18 0 0 1

28 2 22 16 0 0 1

29 2 24 23 0 0 1

30 2 26 27 0 0 1

31 2 31 31 0 0 0

32 2 32 33 0 0 1

33 2 34 37 0 0 1

34 2 35 37 0 0 1

35 2 37 37 0 0 1

36 2 5 3 0 0 0

37 3 0 5 0 0 1

38 3 3 NA 0 0 1

39 3 7 3 0 0 1

40 3 2 4 0 0 0

41 3 5 5 0 0 1

42 3 7 8 0 0 1

43 3 11 10 0 0 1

44 3 5 5 0 0 1

45 3 5 2 0 0 1

46 3 5 6 0 0 0

47 3 3 3 0 0 0

48 3 18 18 0 0 1

49 4 6 8 0 0 0

50 4 5 6 0 0 0

51 4 4 5 0 0 0

52 4 10 9 0 0 1

53 4 8 5 0 0 1

54 4 4 7 0 0 1

55 4 9 9 0 0 1

56 4 6 6 0 0 0

57 4 8 10 0 0 0

58 4 5 6 0 0 0

59 4 7 9 0 0 0

60 4 6 9 0 0 0

61 0 8 6 1 100 1

62 0 7 7 1 100 0

63 0 11 15 1 100 1

64 0 3 6 1 80 1

65 0 5 4 1 80 1

66 0 6 8 1 100 1

67 0 4 4 1 100 0

68 0 9 6 1 80 0

69 0 0 2 1 100 1

70 0 6 7 1 100 1

71 0 5 4 1 60 1

72 0 3 3 1 100 0

73 1 6 5 1 80 0

74 1 6 10 1 100 0

75 1 8 10 1 100 1

76 1 9 15 1 100 0

77 1 10 15 1 100 0

78 1 10 8 1 80 1

79 1 11 15 1 100 0

80 1 11 14 1 80 0

81 1 13 13 1 80 1

82 1 14 17 1 80 1

83 1 17 22 1 100 0

84 1 17 20 1 100 1

85 2 11 16 1 100 1

86 2 11 16 1 100 1

87 2 13 19 1 100 0

88 2 16 21 1 80 0

89 2 24 27 1 100 0

90 2 25 33 1 100 1

91 2 26 30 1 100 1

92 2 30 32 1 100 0

93 2 34 31 1 100 0

94 2 34 35 1 100 1

95 2 35 35 1 100 1

96 2 36 35 1 80 1

97 3 1 3 1 100 1

98 3 7 5 1 100 0

99 3 1 5 1 100 1

100 3 5 2 1 80 0

101 3 2 2 1 80 0

102 3 5 1 1 60 0

103 3 7 7 1 100 1

104 3 3 2 1 80 1

105 3 17 17 1 100 1

106 3 1 1 1 100 0

107 3 9 5 1 100 1

108 4 7 8 1 100 0

109 4 8 8 1 100 0

110 4 6 7 1 80 1

111 4 5 7 1 60 0

112 4 5 9 1 100 1

113 4 8 8 1 100 1

114 4 7 11 1 80 1

115 4 13 13 1 100 0

116 4 6 4 1 60 1

117 4 8 NA 1 60 0

118 4 1 8 1 100 1

> head(dat)

school pre post intervention dose gender

1 0 3 4 0 0 0

```
2   0  7  NA     0  0   0

3   0 12  13     0  0   1

4   0  4  9      0  0   1

5   0  5  8      0  0   0

6   0  2  5      0  0   1

> study <- c("wh")

> out <- matrix(data = NA, nrow = length(study), ncol = 31); colnames(out) <- c("n", "n.sch", "n.t", "n.c",
"icc", "g", "g.lb", "g.ub", "ols", "ols.lb", "ols.ub", "wth", "wth.lb", "wth.ub", "tt", "tt.lb", "tt.ub",
"bs.wth", "bs.wth.lb", "bs.wth.ub", "bs.tt", "bs.tt.lb", "bs.tt.ub", "bt.wth", "bt.wth.lb", "bt.wth.ub",
"bt.tt", "bt.tt.lb", "bt.tt.ub", "pmp.wth", "pmp.tt"); rownames(out) <- study; out

  n n.sch n.t n.c icc g g.lb g.ub ols ols.lb ols.ub wth

wh NA   NA NA NA NA  NA  NA NA   NA   NA NA

  wth.lb wth.ub tt tt.lb tt.ub bs.wth bs.wth.lb bs.wth.ub

wh   NA   NA NA  NA  NA   NA    NA    NA

  bs.tt bs.tt.lb bs.tt.ub bt.wth bt.wth.lb bt.wth.ub bt.tt

wh  NA    NA    NA   NA    NA    NA  NA

  bt.tt.lb bt.tt.ub pmp.wth pmp.tt

wh    NA    NA   NA   NA

> dat$school <- dat$school; length(unique(dat$school)); dat$t <- dat$intervention; dat$pret <-
dat$pre; dat$pret[1:20]

[1] 5

 [1] 3 7 12 4 5 2 8 6 9 6 3 5 4 7 7 9 10 10 11

[20] 12

>

> range(dat$pret, na.rm=T); range(dat$post, na.rm=T)

[1] 0 37
```

```
[1] 1 37

> dat$gain <- dat$post-dat$pret; dat$gain[1:20]

 [1] 1 NA 1 5 3 3 -5 -1 -1 NA 0 2 4 2 0 0 -2 2 -2

[20] -1

> dat$dose

 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0

 [15]  0  0  0  0  0  0  0  0  0  0  0  0  0  0

 [29]  0  0  0  0  0  0  0  0  0  0  0  0  0  0

 [43]  0  0  0  0  0  0  0  0  0  0  0  0  0  0

 [57]  0  0  0  0 100 100 100 80 80 100 100 80 100 100

 [71] 60 100 80 100 100 100 100 80 100 80 80 80 100 100

 [85] 100 100 100 80 100 100 100 100 100 100 100 80 100 100

 [99] 100 80 80 60 100 80 100 100 100 100 100 80 60 100

[113] 100 80 100 60 60 100

> dat.new <- na.omit(dat[c("pret", "post", "gain", "t", "school", "dose")]); dim(dat.new)

[1] 114  6

> out["wh","n"] <- dim(dat.new)[1]; out["wh","n.sch"] <- length(unique(dat.new$school));
out["wh","n.t"] <- nrow(dat.new[dat.new$t==1,]); out["wh","n.c"] <- nrow(dat.new[dat.new$t==0,])

> dat.g <- srtFREQ(post ~ t, intervention="t", data=dat.new)

>

> out["wh","g"] <- dat.g$ES[1,1]; out["wh","g.lb"] <- dat.g$ES[1,2]; out["wh","g.ub"] <- dat.g$ES[1,3]

>

> ## Frequentist OLS

> dat.ols <- srtFREQ(post ~ t + pret + school, intervention="t", data=dat.new)

>
```

```
> out["wh","ols"] <- dat.ols$ES[1,1]; out["wh","ols.lb"] <- dat.ols$ES[1,2]; out["wh","ols.ub"] <- dat.ols$ES[1,3]

>

> ## Frequentist MLM

> dat.mlm <- mstFREQ(post ~ t + pret, random="school", intervention="t", data=dat.new)

Computing profile confidence intervals ...

>

> out["wh","wth"] <- dat.mlm$ES$t[1,1]; out["wh","wth.lb"] <- dat.mlm$ES$t[1,2]; out["wh","wth.ub"] <- dat.mlm$ES$t[1,3]

>

> out["wh","tt"] <- dat.mlm$ES$t[2,1]; out["wh","tt.lb"] <- dat.mlm$ES$t[2,2]; out["wh","tt.ub"] <- dat.mlm$ES$t[2,3]

>

> dat.btmlm <- mstFREQ(post ~ t + pret, random="school", intervention="t", nBoot=1000, data=dat.new)

Computing profile confidence intervals ...

>

> out["wh","bt.wth"] <- dat.btmlm$ES$t[1,1]; out["wh","bt.wth.lb"] <- dat.btmlm$ES$t[1,2]; out["wh","bt.wth.ub"] <- dat.btmlm$ES$t[1,3]

>

> out["wh","bt.tt"] <- dat.btmlm$ES$t[2,1]; out["wh","bt.tt.lb"] <- dat.btmlm$ES$t[2,2]; out["wh","bt.tt.ub"] <- dat.btmlm$ES$t[2,3]

>

> ## Permutation p

> dat.pmp <- mstFREQ(post ~ t + pret, random="school", intervention="t", nPerm=1000, data=dat.new)

Computing profile confidence intervals ...
```

```
>

> perm <- data.frame(dat.pmp$Perm); str(perm)

'data.frame':     1000 obs. of 2 variables:

 $ t1Within: num -0.11 -0.05 -0.07 0.04 -0.44 0.06 -0.11 -0.04 0.04 -0.3 ...

 $ t1Total : num -0.1 -0.05 -0.06 0.04 -0.39 0.05 -0.11 -0.03 0.04 -0.26 ...

> obs.wth <- dat.pmp$ES$t[1,1]; obs.wth

[1] 0.53

> obs.tt <- dat.pmp$ES$t[2,1]; obs.tt

[1] 0.44

> out["wh","pmp.wth"] <- ifelse(mean(perm[,1]>obs.wth)==0, "<.0.001", mean(perm[,1]>obs.wth));
out["wh","pmp.wth"]

[1] 0.003

> out["wh","pmp.tt"] <- ifelse(mean(perm[,2]>obs.tt)==0, "<.0.001", mean(perm[,2]>obs.tt));
out["wh","pmp.tt"]

[1] 0.004

>

> ## Bayesian MLM

> dat.bsmlm <- crtBayes(post ~ t + pret, random="school", intervention="t", nSim=10000,
data=dat.new)

>

> out["wh","icc"] <- round((dat.bsmlm$covParm[[4]]),2); out["wh","icc"]

[1] 0

>

> out["wh","bs.wth"] <- dat.bsmlm$ES[1,1]; out["wh","bs.wth.lb"] <- dat.bsmlm$ES[1,2];
out["wh","bs.wth.ub"] <- dat.bsmlm$ES[1,3]

>
```

```
> out["wh","bs.tt"] <- dat.bsmlm$ES[3,1]; out["wh","bs.tt.lb"] <- dat.bsmlm$ES[3,2]; out["wh","bs.tt.ub"] <- dat.bsmlm$ES[3,3]

> dat.new$dose

 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0

 [15]  0  0  0  0  0  0  0  0  0  0  0  0  0  0

 [29]  0  0  0  0  0  0  0  0  0  0  0  0  0  0

 [43]  0  0  0  0  0  0  0  0  0  0  0  0  0  0

 [57]  0 100 100 100 80 80 100 100 80 100 100 60 100 80

 [71] 100 100 100 100 80 100 80 80 80 100 100 100 100 100

 [85] 80 100 100 100 100 100 100 100 80 100 100 100 80 80

 [99] 60 100 80 100 100 100 100 100 80 60 100 100 80 100

[113] 60 100

> cace.mst <- caceMSTBoot(post ~ pret + t, random = "school", intervention = "t", compliance = "dose", nBoot = 1000, data = dat.new)

> names(cace.mst)

[1] "CACE"   "Compliers"

> cace.mst$CACE

  Compliance  ES  LB  UB

1    P> 0 0.44 0.14 0.75

2    P> 10 0.44 0.14 0.75

3    P> 20 0.44 0.14 0.75

4    P> 30 0.44 0.14 0.75

5    P> 40 0.44 0.14 0.75

6    P> 50 0.44 0.14 0.75

7    P> 60 0.47 0.15 0.80
```

8 P> 70 0.47 0.15 0.80

9 P> 80 0.66 0.22 1.10

10 P> 90 0.66 0.22 1.10

> cace.mst$Compliers

P > 0 P > 10 P > 20 P > 30 P > 40 P > 50 P > 60

pT 1 1 1 1 1 1 0.93

pC 0 0 0 0 0 0 0.00

P=PT-pC 1 1 1 1 1 1 0.93

P > 70 P > 80 P > 90

pT 0.93 0.67 0.67

pC 0.00 0.00 0.00

P=PT-pC 0.93 0.67 0.67

> forest(x=cace.mst$CACE$ES, ci.lb = cace.mst$CACE$LB, ci.ub = cace.mst$CACE$UB, xlab = "CACE ES", xlim = c(-1,2), lwd=1.5, slab = as.character(cace.mst$CACE$Compliance), alim = c(-1,7))

Error: could not find function "forest"

> install.package("metafor")

Error: could not find function "install.package"

> install.packages("metafor")

--- Please select a CRAN mirror for use in this session ---

trying URL 'https://mirrors.ebi.ac.uk/CRAN/bin/macosx/mavericks/contrib/3.2/metafor_1.9-8.tgz'

Content type 'application/x-gzip' length 2176056 bytes (2.1 MB)

==

downloaded 2.1 MB

```
The downloaded binary packages are in
        /var/folders/c1/93d1kdks033_bq9w849tzq5m0000gn/T//RtmpbhnJt0/downloaded_package
s
> library(metafor)
Loading 'metafor' package (version 1.9-8). For an overview
and introduction to the package please type: help(metafor).
> forest(x=cace.mst$CACE$ES, ci.lb = cace.mst$CACE$LB, ci.ub = cace.mst$CACE$UB, xlab = "CACE ES",
xlim = c(-1,2), lwd=1.5, slab = as.character(cace.mst$CACE$Compliance), alim = c(-1,7))
> cace.mst$CACE
  Compliance  ES  LB  UB
1    P> 0 0.44 0.14 0.75
2    P> 10 0.44 0.14 0.75
3    P> 20 0.44 0.14 0.75
4    P> 30 0.44 0.14 0.75
5    P> 40 0.44 0.14 0.75
6    P> 50 0.44 0.14 0.75
7    P> 60 0.47 0.15 0.80
8    P> 70 0.47 0.15 0.80
9    P> 80 0.66 0.22 1.10
10   P> 90 0.66 0.22 1.10
>
```

Values

```
dat.bsmlm              List of 5
  Beta :'data.frame': 3 obs. of 3 variables:
  ..$ Estimate: num [1:3] 0.22 1.34 0.99
  ..$ 95% LB : num [1:3] -0.68 0.36 0.93
  ..$ 95% UB : num [1:3] 1.17 2.32 1.04
  covParm : num [1, 1:4] 0 7.39 7.39 0
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:4] "Schools" "Pupils" "Total" "ICC"
  ES :'data.frame': 3 obs. of 3 variables:
  ..$ Estimate: num [1:3] 5.00e-01 6.44e+31 5.00e-01
  ..$ 95% LB : num [1:3] 1.30e-01 1.53e+14 1.30e-01
  ..$ 95% UB : num [1:3] 8.60e-01 3.64e+32 8.60e-01
  ProbES :'data.frame': 11 obs. of 4 variables:
  ..$ ES : num [1:11] 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 ...
  ..$ t1Within : num [1:11] 1 0.98 0.94 0.86 0.7 0.5 0.29 0.13 0.05 0.02 ...
  ..$ t1Between: num [1:11] 1 1 1 1 1 1 1 1 1 1 ...
  ..$ t1Total : num [1:11] 1 0.98 0.94 0.86 0.7 0.5 0.29 0.13 0.05 0.02 ...
  SchEffects:'data.frame': 5 obs. of 4 variables:
  ..$ Schools : int [1:5] 0 1 2 3 4
  ..$ Estimate: num [1:5] 0 0 0 0 0
  ..$ 95% LB : num [1:5] 0 0 0 0 0
  ..$ 95% UB : num [1:5] 0 0 0 0 0
```

```
dat.btmlm              List of 5
 Beta :'data.frame': 3 obs. of 3 variables:
 ..$ Estimate: num [1:3] 1.13 1.29 0.9
 ..$ 95% LB : num [1:3] -0.53 -0.66 0.83
 ..$ 95% UB : num [1:3] 2.85 3.23 1.01
 covParm : Named num [1:5] 1.7 0.66 5.95 8.31 0.28
 ..- attr(*, "names")= chr [1:5] "Schools" "Intervention:School" "Pupils" "Tota…
 ES :List of 1
 ..$ t1: num [1:2, 1:3] 0.53 0.44 0.18 0.14 0.93 0.76
 .. ..- attr(*, "dimnames")=List of 2
 .. .. ..$ : chr [1:2] "Within" "Total"
 .. .. ..$ : chr [1:3] "Estimate" "95% LB" "95% UB"
 SchEffects:'data.frame': 5 obs. of 2 variables:
 ..$ Schools : num [1:5] 0 1 2 3 4
 ..$ Estimate: num [1:5] -0.24 0.01 0.74 -0.57 0.06
 Bootstrap :'data.frame': 2 obs. of 1000 variables:
 ..$ Estimate : num [1:2] 0.68 0.53
 ..$ Estimate.1 : num [1:2] 0.55 0.45
 ..$ Estimate.2 : num [1:2] 0.88 0.62
 ..$ Estimate.3 : num [1:2] 0.68 0.53
 ..$ Estimate.4 : num [1:2] 0.38 0.32
 ..$ Estimate.5 : num [1:2] 0.52 0.42
 ..$ Estimate.6 : num [1:2] 0.59 0.52
 ..$ Estimate.7 : num [1:2] 0.51 0.42
 ..$ Estimate.8 : num [1:2] 0.31 0.21
 ..$ Estimate.9 : num [1:2] 0.68 0.56
 ..$ Estimate.10 : num [1:2] 0.51 0.4
 ..$ Estimate.11 : num [1:2] 0.33 0.29
 ..$ Estimate.12 : num [1:2] 0.54 0.45
 ..$ Estimate.13 : num [1:2] 0.4 0.35
 ..$ Estimate.14 : num [1:2] 0.48 0.42
 ..$ Estimate.15 : num [1:2] 0.38 0.3
 ..$ Estimate.16 : num [1:2] 0.51 0.35
 ..$ Estimate.17 : num [1:2] 0.55 0.37
 ..$ Estimate.18 : num [1:2] 0.44 0.32
```

```
..$ Estimate.19 : num [1:2] 0.78 0.65
..$ Estimate.20 : num [1:2] 0.63 0.52
..$ Estimate.21 : num [1:2] 0.63 0.46
..$ Estimate.22 : num [1:2] 0.44 0.4
..$ Estimate.23 : num [1:2] 0.72 0.59
..$ Estimate.24 : num [1:2] 0.27 0.22
..$ Estimate.25 : num [1:2] 0.72 0.62
..$ Estimate.26 : num [1:2] 0.73 0.64
..$ Estimate.27 : num [1:2] 0.6 0.48
..$ Estimate.28 : num [1:2] 0.82 0.62
..$ Estimate.29 : num [1:2] 0.52 0.4
..$ Estimate.30 : num [1:2] 0.57 0.41
..$ Estimate.31 : num [1:2] 0.51 0.4
..$ Estimate.32 : num [1:2] 0.66 0.51
..$ Estimate.33 : num [1:2] 0.93 0.67
..$ Estimate.34 : num [1:2] 0.68 0.5
..$ Estimate.35 : num [1:2] 0.66 0.55
..$ Estimate.36 : num [1:2] 0.76 0.68
..$ Estimate.37 : num [1:2] 0.73 0.62
..$ Estimate.38 : num [1:2] 0.54 0.48
..$ Estimate.39 : num [1:2] 0.4 0.28
..$ Estimate.40 : num [1:2] 0.66 0.57
..$ Estimate.41 : num [1:2] 0.45 0.37
..$ Estimate.42 : num [1:2] 0.36 0.26
..$ Estimate.43 : num [1:2] 0.82 0.61
..$ Estimate.44 : num [1:2] 0.64 0.55
..$ Estimate.45 : num [1:2] 0.49 0.44
..$ Estimate.46 : num [1:2] 0.4 0.3
..$ Estimate.47 : num [1:2] 0.74 0.58
..$ Estimate.48 : num [1:2] 0.46 0.35
..$ Estimate.49 : num [1:2] 0.83 0.66
..$ Estimate.50 : num [1:2] 0.67 0.54
..$ Estimate.51 : num [1:2] 0.62 0.52
..$ Estimate.52 : num [1:2] 0.64 0.52

..$ Estimate.53 : num [1:2] 0.57 0.42
..$ Estimate.54 : num [1:2] 0.38 0.28
..$ Estimate.55 : num [1:2] 0.28 0.19
..$ Estimate.56 : num [1:2] 0.53 0.48
..$ Estimate.57 : num [1:2] 0.89 0.75
..$ Estimate.58 : num [1:2] 0.68 0.55
..$ Estimate.59 : num [1:2] 0.55 0.4
..$ Estimate.60 : num [1:2] 0.61 0.49
..$ Estimate.61 : num [1:2] 0.2 0.19
..$ Estimate.62 : num [1:2] 0.61 0.56
..$ Estimate.63 : num [1:2] 0.43 0.37
..$ Estimate.64 : num [1:2] 0.81 0.76
..$ Estimate.65 : num [1:2] 0.49 0.38
..$ Estimate.66 : num [1:2] 0.81 0.72
..$ Estimate.67 : num [1:2] 0.6 0.46
..$ Estimate.68 : num [1:2] 0.61 0.5
..$ Estimate.69 : num [1:2] 0.83 0.66
..$ Estimate.70 : num [1:2] 0.74 0.57
..$ Estimate.71 : num [1:2] 0.23 0.19
..$ Estimate.72 : num [1:2] 0.6 0.49
..$ Estimate.73 : num [1:2] 0.44 0.32
..$ Estimate.74 : num [1:2] 0.36 0.32
..$ Estimate.75 : num [1:2] 0.6 0.5
..$ Estimate.76 : num [1:2] 0.74 0.6
..$ Estimate.77 : num [1:2] 0.38 0.27
..$ Estimate.78 : num [1:2] 0.43 0.4
..$ Estimate.79 : num [1:2] 0.67 0.65
..$ Estimate.80 : num [1:2] 0.58 0.47
..$ Estimate.81 : num [1:2] 0.61 0.52
..$ Estimate.82 : num [1:2] 0.05 0.04
..$ Estimate.83 : num [1:2] 0.71 0.6
```

```
dat.g                  List of 3
  Beta :'data.frame': 2 obs. of 3 variables:
  ..$ Estimate: num [1:2] 10.98 1.28
  ..$ 95% LB : num [1:2] 8.53 -2.19
  ..$ 95% UB : num [1:2] 13.44 4.75
  ES :'data.frame': 1 obs. of 3 variables:
  ..$ Estimate: num 0.14
  ..$ 95% LB : num -0.23
  ..$ 95% UB : num 0.5
  sigma2: num 87.5
dat.mlm                List of 4
  Beta :'data.frame': 3 obs. of 3 variables:
  ..$ Estimate: num [1:3] 1.13 1.29 0.9
  ..$ 95% LB : num [1:3] -0.53 -0.66 0.83
  ..$ 95% UB : num [1:3] 2.85 3.23 1.01
  covParm : Named num [1:5] 1.7 0.66 5.95 8.31 0.28
  ..- attr(*, "names")= chr [1:5] "Schools" "Intervention:School" "Pupils" "Tota…
  ES :List of 1
  ..$ t1:'data.frame': 2 obs. of 3 variables:
  .. ..$ Estimate: num [1:2] 0.53 0.44
  .. ..$ 95% LB : num [1:2] -0.34 -0.3
  .. ..$ 95% UB : num [1:2] 1.4 1.17
  SchEffects:'data.frame': 5 obs. of 2 variables:
  ..$ Schools : num [1:5] 0 1 2 3 4
  ..$ Estimate: num [1:5] -0.24 0.01 0.74 -0.57 0.06
dat.ols                List of 3
  Beta :'data.frame': 4 obs. of 3 variables:
  ..$ Estimate: num [1:4] 0.12 1.34 0.99 0.05
  ..$ 95% LB : num [1:4] -1.09 0.33 0.93 -0.31
  ..$ 95% UB : num [1:4] 1.34 2.35 1.04 0.41
  ES :'data.frame': 1 obs. of 3 variables:
  ..$ Estimate: num 0.49
  ..$ 95% LB : num 0.12
  ..$ 95% UB : num 0.86
  sigma2: num 7.35
dat.pmp                List of 5
  Beta :'data.frame': 3 obs. of 3 variables:
  ..$ Estimate: num [1:3] 1.13 1.29 0.9
  ..$ 95% LB : num [1:3] -0.53 -0.66 0.83
  ..$ 95% UB : num [1:3] 2.85 3.23 1.01
  covParm : Named num [1:5] 1.7 0.66 5.95 8.31 0.28
  ..- attr(*, "names")= chr [1:5] "Schools" "Intervention:School" "Pupils" "Tota…
  ES :List of 1
  ..$ t1:'data.frame': 2 obs. of 3 variables:
  .. ..$ Estimate: num [1:2] 0.53 0.44
  .. ..$ 95% LB : num [1:2] -0.34 -0.3
  .. ..$ 95% UB : num [1:2] 1.4 1.17
  SchEffects:'data.frame': 5 obs. of 2 variables:
  ..$ Schools : num [1:5] 0 1 2 3 4
  ..$ Estimate: num [1:5] -0.24 0.01 0.74 -0.57 0.06
  Perm :'data.frame': 1000 obs. of 2 variables:
  ..$ t1Within: num [1:1000] -0.11 -0.05 -0.07 0.04 -0.44 0.06 -0.11 -0.04 0.04 …
  ..$ t1Total : num [1:1000] -0.1 -0.05 -0.06 0.04 -0.39 0.05 -0.11 -0.03 0.04 -…
new.packages           character (empty)
obs.tt                 0.44
obs.wth                0.53
required.packages      chr [1:10] "ggplot2" "mvtnorm" "xtable" "psych" "metafor" …
study                  "wh"
```

Appendix 11: R Code Study One Additional Analysis

```
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> load("C:/Users/wdhar/OneDrive/Desktop/PhD drafts/Chapter 5 Trial 1 Online peer tuition/R data/ANALYSIS.RData")
> #rm(list = ls())
> setwd("C:/Users/wdhar/OneDrive/Desktop/PhD drafts/Chapter 5 Trial 1 Online peer tuition/R data")
> load()
> library(nlme)
> wdat <- read.csv("wdata4.csv", header=T); wdat
   i..Pupil.Code Gender School Group.Size Role PP pret post
1              1      M      0      01:04   PT  0   15   17
2              2      F      0      01:01   PT  0    7   17
3              3      F      0      01:02   PT  0   11   17
4              4      F      0      01:01   PT  0   17   17
5              5      M      0      01:02   PT  0   10   15
6              6      M      0      01:01   PT  0   13    9
7              7      M      0      01:01   PT  1    9   14
8              9      M      0      01:02   PT  1   13   15
9             10      M      0      01:01   PT  0   14   15
10            11      M      0      01:04   PT  0   17   16
11            12      F      0      01:01   PT  1    8   15
12            13      M      0      01:01   PT  0   17   16
13            14      M      0      01:01   PT  0   14   17
14            15      M      0      01:01   PT  0   13   15
15            16      M      0      01:01   PT  0   15   17
16            17      M      0      01:01   PT  1    5   15
17            18      F      0      01:02   PT  0   10   16
18            19      F      0      01:04   PT  0   15   17
19            20      M      0      01:04   PT  0   11   16
20            21      M      0      01:02   PT  0    8   15
21            22      M      0      01:01   PT  1   11   11
22            23      F      1      01:04    P  0    3    6
23            24      M      1      01:01    P  0    9   17
24            25      M      1      01:02    P  1    3   10
25            26      M      1      01:01    P  0    3   15
26            27      M      1      01:04    P  0    2   14
27            28      F      1      01:01    P  0    2   12
28            29      F      1      01:04    P  0    4   17
29            30      F      1      01:02    P  0    4   12
30            31      F      1      01:02    P  1    2   12
31            32      F      1      01:02    P  0    5   15
32            33      F      1      01:01    P  0    4   11
33            34      M      2      01:04    P  0    2    4
34            35      M      2      01:02    P  0    7   16
35            36      M      2      01:01    P  1    2    7
36            37      M      2      01:02    P  0    4   12
37            38      F      2      01:04    P  0    3    7
38            39      M      2      01:02    P  1    1    3
39            40      F      2      01:04    P  1    7   10
40            41      M      2      01:04    P  0    2    7
41            42      M      2      01:01    P  0    2    9
42            44      F      2      01:01    P  0    4   11
```

```
43            45      M      2      01:04    P  0    3   10
44            46      M      2      01:01    P  1   10   13
45            47      F      2      01:04    P  0    5    7
46            48      F      2      01:01    P  0    7   11
47            49      F      2      01:04    P  0    1    7
48            50      F      2      01:02    P  0    2    7
49            51      F      2      01:01    P  0    1   14
50            52      F      2      01:01    P  1    2    8
51            53      M      2      01:02    P  0    2   10
52            54      F      2      01:04    P  0    2    9
53            55      M      2      01:01    P  0    3    8
54            56      F      2      01:02    P  0    3    6
55            57      F      2      01:01    P  1    0    9
56            58      M      2      01:02    P  0    8   13
57            59      F      3      00:00   PT  0   11    9
58            60      M      3      00:00   PT  0    8   12
59            62      M      3      00:00   PT  0    7   11
60            63      F      3      00:00   PT  0   15   11
61            64      F      3      00:00   PT  0   12   10
62            65      M      3      00:00   PT  0   13   12
63            66      F      3      00:00   PT  0   12   15
64            68      F      4      00:00    P  1    3    1
65            69      M      4      00:00    P  0    4    2
66            70      F      4      00:00    P  0    7    9
67            71      F      4      00:00    P  0    4    6
68            72      M      4      00:00    P  0    3    8
69            73      M      4      00:00    P  1    1    1
70            74      M      4      00:00    P  0    0    3
71            75      M      4      00:00    P  1    0    3
72            76      M      4      00:00    P  0    1    1
73            77      M      4      00:00    P  0    0    4
74            78      F      4      00:00    P  0    5    7
75            79      M      4      00:00    P  0    2    2
> wdat$female <- NA; wdat$female[wdat$Gender %in% "M"] <- 0; wdat$female[wdat$Gender %in% "F"] <- 1; wdat$female
 [1] 0 1 1 1 0 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 1 1 1 1 1 1 0 0 0 0 1 0 1 0 0 1 0 0 1 1 1 1 1 1 0 1 0 1 1 0 1 0
[59] 0 1 1 0 1 1 0 1 1 0 0 0 0 0 0 1 0
> wdat$t <- NA; wdat$t[wdat$Group.Size %in% "01:04"] <- 3; wdat$t[wdat$Group.Size %in% "01:02"] <- 2; wdat$t[wdat$Group.Size
%in% "01:01"] <-1; wdat$t
 [1]  3  1  2  1  2  1  1  2  1  3  1  1  1  1  1  1  2  3  3  2  1  3  1  2  1  3  1  3  2  2  2  1  3  2  1  2  3  2  3
[40]  3  1  1  3  1  3  1  3  2  1  1  2  3  1  2  1  2 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
> wdat$t[wdat$Group.Size %in% "00:00"] <- 0;
> wdat$post <- wdat$post; wdat$pret <- wdat$pret
> head(wdat)
  i..Pupil.Code Gender School Group.Size Role PP pret post female t
1             1      M      0      01:04   PT  0   15   17      0 3
2             2      F      0      01:01   PT  0    7   17      1 1
3             3      F      0      01:02   PT  0   11   17      1 2
4             4      F      0      01:01   PT  0   17   17      1 1
5             5      M      0      01:02   PT  0   10   15      0 2
6             6      M      0      01:01   PT  0   13    9      0 1

> wdat <- na.omit(wdat[c("post", "t", "pret", "School", "female")]); dim(wdat)
[1] 75  5
> head(wdat)
  post t pret School female
1   17 3   15      0      0
2   17 1    7      0      1
3   17 2   11      0      1
4   17 1   17      0      1
5   15 2   10      0      0
6    9 1   13      0      0
> wd.0to2 <- subset(wdat, wdat$t==2 | wdat$t==0); dim(wd.0to2)
[1] 35  5
> wd.0to1 <- subset(wdat, wdat$t==1 | wdat$t==0); dim(wd.0to1)
[1] 44  5
> wd.0to3 <- subset(wdat, wdat$t==3 | wdat$t==0); dim(wd.0to3)
[1] 34  5
> wd.0to1$t
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> wd.0to1$post
```

```
 [1] 17 17  9 14 15 15 16 17 15 17 15 11 17 15 12 11  7  9 11 13 11 14  8  8  9  9 12 11 11 10 12 15  1  2  9  6  8  1  3
[40]  3  1  4  7  2
> efs2(wd.0to2$post, wd.0to2$t)
     cd cd.lb cd.ub     g g.lb g.ub    d d.lb d.ub
1 1.27  0.54     2 1.24 0.53 1.95 1.21 0.48 1.94
> efs(wd.0to1$post, wd.0to1$t)
     cd cd.lb cd.ub     g g.lb g.ub    d d.lb d.ub
1 1.62  0.93   2.3 1.59 0.92 2.26 1.39  0.7 2.08
> ancovaES2(wd.0to2$post, wd.0to2$t, wd.0to2$pret)
     cd cd.lb cd.ub     g  g.lb g.ub beta sd.pool
1 0.32 -0.35  0.99 0.31 -0.34 0.97  0.8     2.48
> ancovaES(wd.0to1$post, wd.0to1$t, wd.0to1$pret)
     cd cd.lb cd.ub     g g.lb g.ub beta sd.pool
1 1.88  1.17   2.6 1.85 1.15 2.55 5.19     2.76
> #control to 1:2
> lme <- lme(post ~ t + pret, random = ~ 1|School, data=wd.0to2, na.action=na.exclude); summary(lme)
Linear mixed-effects model fit by REML
 Data: wd.0to2
       AIC      BIC    logLik
  172.1096 179.4383 -81.05482

Random effects:
 Formula: ~1 | School
        (Intercept) Residual
StdDev:   0.8783973  2.41277

Fixed effects: post ~ t + pret
                Value Std.Error DF  t-value p-value
(Intercept) 2.3894076 1.1283147 29 2.117678  0.0429
t           2.6941286 0.5779579  3 4.661462  0.0186
pret        0.7666336 0.1197336 29 6.402829  0.0000
 Correlation:
     (Intr) t
t    -0.569
pret -0.668  0.043

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.0658661 -0.7760898  0.2553271  0.6628961  1.4712246


Number of Observations: 35
Number of Groups: 5
> beta.0to2 <- data.frame(intervals(lme, which="fixed")$fixed)[2,2]; beta.0to2
[1] 2.694129
> var.b.0to2 <- as.numeric(getVarCov(lme)); var.b.0to2
[1] 0.7715819
> var.w.0to2 <- lme$sigma^2; var.w.0to2
[1] 5.821459
> var.tt.0to2 <- var.w.0to2 + var.b.0to2; var.tt.0to2
[1] 6.593041
> icc.0to2 <- round(as.numeric(getVarCov(lme))/(as.numeric(getVarCov(lme))+lme$sigma^2),2); icc.0to2
[1] 0.12
> length(unique(wd.0to2$School))
[1] 5
> tt.0to2 <- g.total2(var.tt.0to2, beta.0to2, icc.0to2, wd.0to2$t, wd.0to2$School); tt.0to2
     g  LB   UB
1 1.02 0.1 1.95
> wth.0to2 <- g.within2(var.w.0to2, beta.0to2, icc.0to2, wd.0to2$t, wd.0to2$School); wth.0to2
     g   LB   UB
1 1.12 0.12 2.11
> btw.0to2 <- g.between2(var.b.0to2, beta.0to2, icc.0to2, wd.0to2$t, wd.0to2$School); btw.0to2
     g LB    UB
1 4.59 -3 12.18
> #control to 1:1
> lme <- lme(post ~ t + pret, random = ~ 1|School, data=wd.0to1, na.action=na.exclude); summary(lme)
Linear mixed-effects model fit by REML
 Data: wd.0to1
       AIC      BIC    logLik
  216.8048 225.3726 -103.4024
```

```
Random effects:
 Formula: ~1 | School
        (Intercept) Residual
StdDev:    2.209159 2.408902

Fixed effects: post ~ t + pret
               Value Std.Error DF  t-value p-value
(Intercept) 5.228038 1.8233567 38 2.867260  0.0067
t           5.258698 2.1608172  3 2.433662  0.0930
pret        0.348071 0.1118143 38 3.112939  0.0035
 Correlation:
     (Intr) t
t    -0.700
pret -0.410 -0.004

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.57066992 -0.72866317 -0.03949685  0.61502824  1.93205575

Number of Observations: 44
Number of Groups: 5
> beta.0to1 <- data.frame(intervals(lme, which="fixed")$fixed)[2,2]; beta.0to1
[1] 5.258698
> var.b.0to1 <- as.numeric(getVarCov(lme)); var.b.0to1
[1] 4.880382
> var.w.0to1 <- lme$sigma^2; var.w.0to1
[1] 5.80281
> var.tt.0to1 <- var.w.0to1 + var.b.0to1; var.tt.0to1
[1] 10.68319
> icc.0to1 <- round(as.numeric(getVarCov(lme))/(as.numeric(getVarCov(lme))+lme$sigma^2),2); icc.0to1
[1] 0.46
> length(unique(wd.0to1$School))
[1] 5
> tt.0to1 <- g.total(var.tt.0to1, beta.0to1, icc.0to1, wd.0to1$t, wd.0to1$School); tt.0to1
     g   LB   UB
1 1.44 0.03 2.86
> wth.0to1 <- g.within(var.w.0to1, beta.0to1, icc.0to1, wd.0to1$t, wd.0to1$School); wth.0to1
     g   LB   UB
1 2.18 0.29 4.08
> btw.0to1 <- g.between1(var.b.0to1, beta.0to1, icc.0to1, wd.0to1$t, wd.0to1$School); btw.0to1
     g    LB   UB
1 2.56 -1.86 6.99
> #control to 1:4
> wd.0to3$t[wd.0to3$t==3]<-1
> lme <- lme(post ~ t + pret, random = ~ 1|School, data=wd.0to3, na.action=na.exclude); summary(lme)
Linear mixed-effects model fit by REML
 Data: wd.0to3
       AIC      BIC    logLik
  171.6846 178.8545 -80.84228

Random effects:
 Formula: ~1 | School
        (Intercept) Residual
StdDev:    2.220648 2.523881

Fixed effects: post ~ t + pret
               Value Std.Error DF  t-value p-value
(Intercept) 3.926815 1.9608300 28 2.002629  0.0550
t           4.329283 2.2253857  3 1.945408  0.1469
pret        0.542175 0.1514323 28 3.580312  0.0013
 Correlation:
     (Intr) t
t    -0.643
pret -0.516 -0.008

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.2171046 -0.7600262  0.1010150  0.6517925  1.9264449


Number of Observations: 34
Number of Groups: 5
> beta.0to3 <- data.frame(intervals(lme, which="fixed")$fixed)[2,2]; beta.0to3
[1] 4.329283
> var.b.0to3 <- as.numeric(getVarCov(lme)); var.b.0to3
[1] 4.931279
> var.w.0to3 <- lme$sigma^2; var.w.0to3
[1] 6.369974
> var.tt.0to3 <- var.w.0to3 + var.b.0to3; var.tt.0to3
[1] 11.30125
> icc.0to3 <- round(as.numeric(getVarCov(lme))/(as.numeric(getVarCov(lme))+lme$sigma^2),2); icc.0to3
[1] 0.44
> length(unique(wd.0to3$School))
[1] 5
> tt.0to3 <- g.total(var.tt.0to3, beta.0to3, icc.0to3, wd.0to3$t, wd.0to3$School); tt.0to3
     g    LB   UB
1 1.16 -0.24 2.56
> wth.0to3 <- g.within(var.w.0to3, beta.0to3, icc.0to3, wd.0to3$t, wd.0to3$School); wth.0to3
     g    LB   UB
1 1.72 -0.15 3.58
> btw.0to3 <- g.between1(var.b.0to3, beta.0to3, icc.0to3, wd.0to3$t, wd.0to3$School); btw.0to3
     g    LB   UB
1 2.18 -2.21 6.57
> save.image("ANALYSIS.RData")
>

>
```

Appendix 12: Fidelity measures for Study One and Study Two

Study One

The fidelity measure consisted of three component scores relating to (a) how well the technology worked (b) the tutor attitudes to learning (c) content delivered in the 30-minute lesson. Each component was rated between 1 and 3, with 1 corresponding to poor, 2 corresponding to partially worked and 3 corresponding to worked as planned. Therefore, each lesson could receive a fidelity score between 3 and 9.

The average school completed 7 lessons per week concurrently in a 30-minute timescale, with a single score for each component made based on the judgement of the researcher. The average score for each component was calculated from the 6 lessons observed to allow an overall fidelity score for each secondary school delivering the online peer tutoring.

Primary schools were not included as the fidelity score had been based on the researcher observations and it was only possible to be present at one school at a time.

School	Component A Technology	Component B Attitude	Component C Lesson content	Fidelity score
0	2.5	3	2.3	7.8
5	2.7	3	2.5	8.2
9	2.8	3	2.3	8.1

Study Two

A similar methodology to the fidelity score was used for the researcher observations for the five schools involved in the online small group teaching. As one school involved in the programme was located outside of the North East, this was excluded from the researcher fidelity score. However, a second fidelity measure using teacher evaluation forms provided by the online learning company provided a fidelity score for each school. The teacher evaluation forms scored each participating pupil for achievement, effort and behaviour for learning from a score of 1 – 5, with 1 corresponding to poor and 5 corresponding to excellent. The average score for each school was calculated for each

measure to provide a fidelity measure for each metric for participating schools. Consequently, the fidelity measure used in this thesis was the teacher lesson feedback form for Study 2.

The following tables show the fidelity scores for both the researcher observations and teacher evaluations for participating schools.

Online small group teaching researcher observations fidelity scores

School	Component A Technology	Component B Attitude	Component C Lesson content	Fidelity score
0	2.7	2.5	2.5	7.7
1	2.8	2.6	2.5	7.9
2	2.7	3	2.7	8.4
3	2.8	2.1	1.9	6.8
4	NA	NA	NA	NA

Online small group teaching teacher evaluation fidelity scores

School	Achievement	Effort	Behaviour for learning
0	3.68	4.12	4.18
1	3.94	4.48	4.74
2	3.9	4.42	4.86
3	3.3	3.66	3.22
4	3.72	3.86	3.76